Decision Support Systems

MEMF: Multi-Entity Multimodal Fusion Framework for Sales Prediction in Live Streaming Commerce

--Manuscript Draft--

Manuscript Number:	DECSUP-D-23-01630
Article Type:	Research Paper
Keywords:	live streaming commerce; sales prediction; multimodal fusion; multi-entity fusion
Abstract:	Live streaming commerce enriched with multimodal information has emerged as a dominant trend in the field of e-commerce. However, analyzing and predicting sales for live-streaming commerce with these data remains a significant challenge. This study intends to propose a framework that fuses multi-entity and multi-modal information for accurate sales predictions. We first fuse multimodal information under each entity according to the anchor, commodity, and live streaming room. On the basis of this foundation, we further integrate video and audio features to obtain a comprehensive feature representation. Moreover, to improve the representation of commodity entity, the Multimodal QuadTransformer is introduced for feature extraction. Experiments are conducted on a real-world dataset in Taobao Live. The results show the framework’s outstanding performance in sales predictions. Compared to existing methods, our framework not only considers more information modalities but also emphasizes entity-level information fusion, demonstrating its practicality in live-streaming e-commerce.

MEMF: Multi-Entity Multimodal Fusion Framework for Sales Prediction in Live Streaming Commerce

Abstract: Live streaming commerce enriched with multimodal information has emerged as a dominant trend in the field of e-commerce. However, analyzing and predicting sales for live-streaming commerce with these data remains a significant challenge. This study intends to propose a framework that fuses multi-entity and multi-modal information for accurate sales predictions. We first fuse multimodal information under each entity according to the anchor, commodity, and live streaming room. On the basis of this foundation, we further integrate video and audio features to obtain a comprehensive feature representation. Moreover, to improve the representation of commodity entity, the Multimodal QuadTransformer is introduced for feature extraction. Experiments are conducted on a real-world dataset in Taobao Live. The results show the framework's outstanding performance in sales predictions. Compared to existing methods, our framework not only considers more information modalities but also emphasizes entity-level information fusion, demonstrating its practicality in live-streaming e-commerce.

Keywords: live streaming commerce; sales prediction; multimodal fusion; multi-entity fusion

1. Introduction

In recent years, live streaming commerce has gradually become a mainstream trend in e-commerce, significantly boosting its development [1]. Compared to traditional e-commerce, live streaming commerce revolutionizes the communication methods between consumers and merchants. In the live streaming environment, the anchors create real-time content through the platform to provide consumers with richer and more intuitive information [2,3] in addition to text, images, and numbers, enhancing the interaction between them and improving the shopping experience of consumers [4,5]. As more merchants and anchors enter the live streaming commerce, competition is becoming increasingly fierce [6]. In such a competitive environment, optimizing the strategies and ensuring maximum economic returns for each live streaming have become central concerns. At this point, accurate prediction of sales is particularly important. With accurate sales prediction, merchants and anchors can not only adequately prepare for live streaming and ensure optimal inventory allocation, but also devise more effective marketing strategies, leading to enhanced revenue [7].

Contemporary live streaming platforms offer information that extends well beyond traditional numerical and textual data, encompassing a rich array of acoustic, visual, and other multimodal data. While these multimodal data present great potential in predicting live streaming sales, challenges still persist in exploring and harnessing complementary information across different modalities [8,9]. One significant challenge is the effective fusion of this multimodal information. Recently, studies on live streaming's multimodal information have emerged, focusing on areas like commodity return prediction [4], gift-giving behavior [10], and traffic prediction [11]. These studies, however, have two major limitations due to the lack of consideration of the characteristics of live streaming. First, most research predominantly emphasizes numerical and textual data. Considering live streaming involves text, numbers, images, audio and video, existing fusion methods might not fully adapt to the complexity and dynamism of its environment. Second, almost all fusion strategies use the same approach, processing similar modalities together before fusing them. However, such strategies appear insufficient in effectively addressing the tasks of live streaming commerce. Within the context of live streaming, the multimodal information transmitted to consumers by the platform primarily originates from three entities: The anchor, the commodity being showcased, and the live streaming room [12]. As we all know, the correlation between different modal information in the same entity is actually much greater than that of the same modal information in different entities. These strategies undoubtedly ignore the complex correlation between different modal information of the same entity in the live streaming environment, which affect the performance of downstream tasks.

To this end, the goal of this study is to propose an improved multimodal fusion framework for accurate sales prediction during live streaming. On the one hand, we ensure a comprehensive incorporation of diverse modalities, including text, images, numerical data, video, and audio in our framework. On the other hand, recognizing the information disparities in different entities, we incorporate a multi-entity fusion approach on top of the multimodal fusion to facilitate more accurate sales prediction. Specifically, we categorize the multimodal information into the three entities. Each entity consists of numerical, textual, and image-based modalities, and uses Transformer architecture to fuse the cross-modal information. Subsequently, the feature representations from these three entities are fused at the entity level with the video and audio features to obtain a more accurate multimodal feature representation influencing live streaming sales.

The primary contributions of this study are manifold. Firstly, we design a multi-entity multimodal information fusion framework for live streaming sales prediction, and carry out comparative experiments with a range of existing multimodal fusion methods. This comparison emphasizes the importance of introducing multi-entity fusion into our method. Secondly, we propose the Multimodal QuadTransformer method to extract multimodal information of the commodity entity in live streaming, which takes intra-commodity modal features and inter-commodity feature relations into account comprehensively. Thirdly, we evaluated the influence of the different entities' information on sales prediction performance, and conclude that information of commodity entity has the most significant impact. Finally, we test the robustness of the framework, specifically evaluating the effects of image and text modal feature extraction methods on sales prediction performance.

The remainder of this paper is structured as follows. Section 2 reviews related work on live streaming commerce, sales prediction, and multimodal fusion. Section 3 describes the proposed framework and the computational methodologies employed. Section 4 presents the experimental setup, results, and discussion. Section 5 concludes the research and discusses the future work.

2. Related work

2.1 Live streaming commerce

Live streaming commerce, as a novel business model, crafts a more intuitive and engaging shopping environment for consumers by integrating the shopping experience of traditional e-commerce with real-time video interaction [13]. This not only provides a new sales avenue for merchants but also deepens their interaction with consumers [1], enhancing the shopping experience and consequently bolstering customer loyalty [3]. In addition, live streaming commerce provides a more realistic and multidimensional display of products, greatly reducing the inherent uncertainty of consumers' online shopping [14].

As an integral component of live streaming commerce, streaming platforms not only display basic commodity details like price and description but also enrich the content with multimodal information from anchors, such as audio and video. This information can be broadly categorized into three entities, the anchor, the commodity, and the live streaming room [12]. This information plays an important role throughout the streaming process, exerting significant influence on the success of the live streaming and related business decisions. Previous empirical studies have shown that the influence of anchors has a significant impact on audience engagement and information transmission [15]. Meanwhile, the showcasing and recommendations of commodity are particularly critical in live streaming commerce [5,16]. They not only dictate the commercial value of the live streaming but also sway audience’s purchase decisions. Concerning the live streaming room, a few studies have highlighted the ambiance, content vividness, and other attributes as critical determinants of a live streaming’s success [17]. With the ascent of live streaming commerce, research in this domain is burgeoning. However, most existing studies predominantly employ empirical methods to investigate user engagement behaviors, purchase intention determinants, and anchors’ influence. Yet, research on using the multimodal information offered by streaming platforms for sales prediction still remain scant. This study delves into this aspect, specifically leveraging the multimodal information provided by live streaming platforms for sales predictions.

2.2 Sales prediction

Accurate sales prediction can assist enterprises in more precisely managing inventory, pricing strategies, marketing promotions, and ultimately maximizing profits [7]. Traditional sales prediction methods primarily rely on historical sales data, employing statistical and economic methods to predict future sales. These methods, such as the ARIMA [18] and ES [19], hold certain predictive accuracies for stable market conditions and specific data patterns. However, in the realm of e-commerce especially live streaming commerce, these methods may no longer be applicable, because multimodal information in e-commerce can influence sales, complicating the prediction process.

In recent years, sales prediction methods have undergone profound transformations. The emergence of deep learning technology, such as CNN [20], RNN [21], and Transformer [22], provides a new paradigm for processing complex, nonlinear and high-dimensional data. This evolution enables sales prediction to use more diverse and enriched dataset and achieve greater predictive accuracy. For instance, Joseph et al. [20] employ CNN to extract feature vectors and Bi-LSTM to learn temporal patterns, subsequently effectively predicting commodity demand. Additionally, some studies also try to integrate multiple data sources, such as social media data [23] and user comments [24], to improve the accuracy of prediction. Yet, the majority of existing research primarily focuses on processing text or numerical modality data, ignoring the rich acoustic, visual and other information in the live streaming environment.

Within the context of live streaming commerce, sales prediction encounters distinct and complex challenges. Firstly, due to the real-time interaction of live streaming commerce [2], sales data might be influenced by various factors like the anchors' performance, audience feedback, and the showcasing of commodity. This renders sales data highly volatile and nonlinear. Secondly, live streaming platform provides a large number of multimodal information [4], including text, image, numerical data, audio, and video. There may be complex correlations and influences between these information. Effectively extracting and fusing meaningful features from the information becomes an important concern for sales prediction. To address this issue, we propose a multi-entity multimodal information fusion framework to predict sales in live streaming commerce, which takes full advantage of different modal information.

2.3 Multimodal fusion

Multimodal fusion refers to extracting information from multiple different modalities and integrating the information into a unified framework to improve the performance of downstream tasks, such as recognition [25], classification [26] and retrieval [27]. Early research has explored various strategies for multimodal fusion [28], including kernel-based methods, graphical models, and neural networks. However, with the rapid development of computer technology, the application of deep neural networks, especially CNN [29], RNN [30], and Transformer structures [31], has become increasingly prevalent in the field of multimodal fusion. This method provides powerful representation ability for the model, especially when dealing with multimodal data. Current mainstream fusion strategies can be roughly categorized into bilinear operations and attention mechanisms. Bilinear operations mainly rely on matrix or tensor mathematical operations to achieve interaction between modalities, such as MFB [32], TFN [33], and MFH [34]. Attention mechanisms provides a dynamic weight allocation strategy for multimodal fusion, allowing the model to determine the importance of one modality based on the content of another. Self-attention [35] and cross-modal attention [36] in the Transformer structure are typical applications of this strategy.

At present, some researchers have integrated multimodal information of live streaming commerce through deep neural network method, and then completed downstream tasks, such as live streaming recommendation [37], emotion analysis [38], user gift giving behavior prediction [10] and traffic prediction [11]. Most of these studies centralize and uniformly process information of the same modality to obtain its feature representation when fusing multimodal information. In fact, for the same modality information, its semantics on different entities often have significant differences. Simply processing the information through a unified model may result in the loss of some information, which will affect the performance of analysis and prediction. Therefore, considering the differences of modalities on different entities, we propose to introduce the entity level fusion method based on the traditional multimodal fusion, in order to achieve a more accurate sales prediction.

3. The MEMF framework

In this section, we propose the **M**ulti-**E**ntity **M**ultimodal **F**usion framework (MEMF) for sales prediction in live streaming commerce. As depicted in Figure 1, the MEMF consists of four components, including data collection, feature extraction, feature fusion, and sales prediction. In the following, we will next introduce each component in detail.

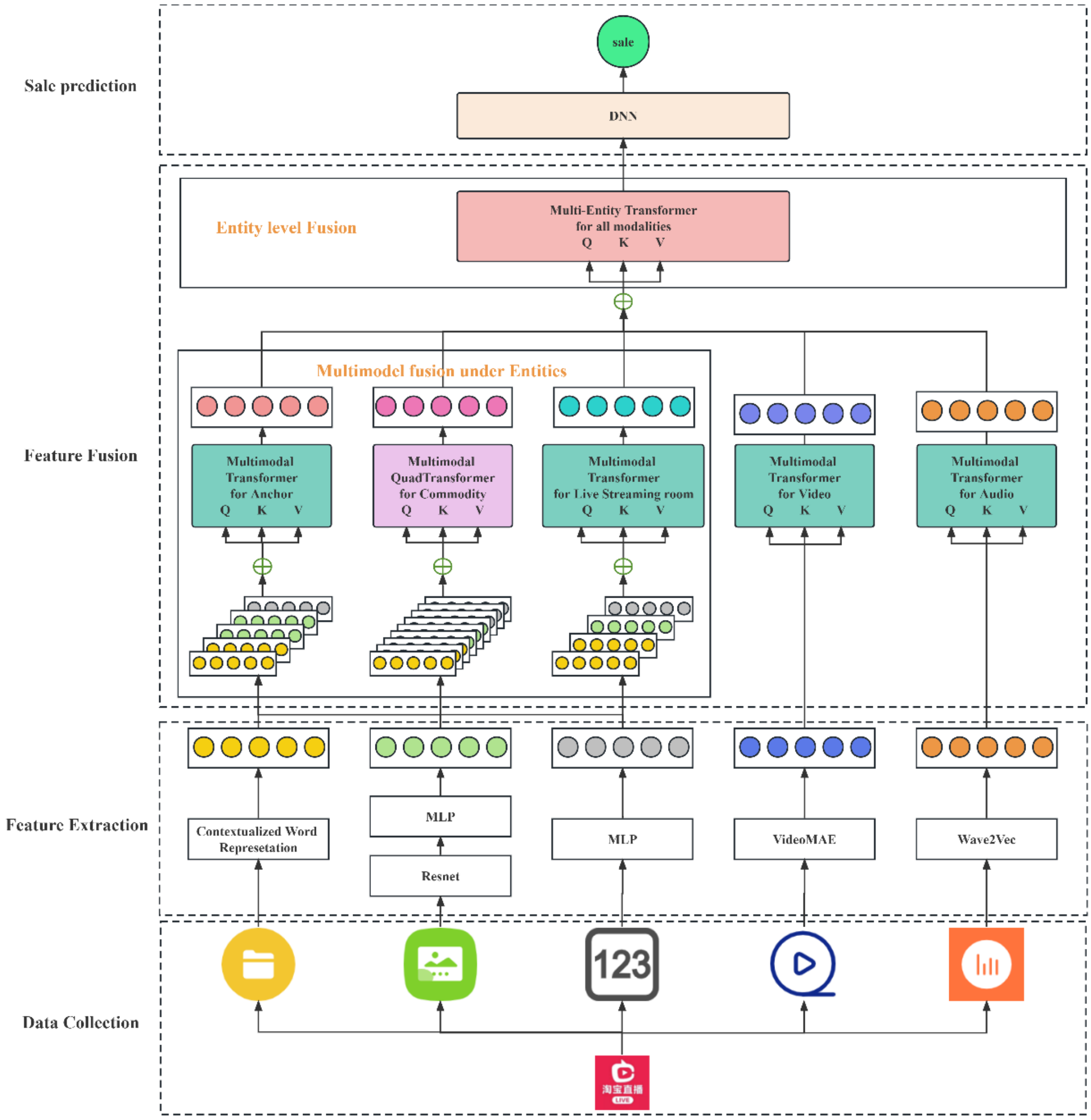
Sale prediction
sale
DNN
Entity level Fusion
Multi-Entity Transformer for all modalities
Q K V
Feature Fusion
Multimodel fusion under Entities
Multimodal Transformer for Anchor
Q K V
Multimodal QuadTransformer for Commodity
Q K V
Multimodal Transformer for Live Streaming room
Q K V
Multimodal Transformer for Video
Q K V
Multimodal Transformer for Audio
Q K V
Feature Extraction
Contextualized Word Represetation
MLP
Resnet
MLP
VideoMAE
Wave2Vec
Data Collection
123
淘宝直播
LIVE

Fig.1. Proposed framework for sales prediction

3.1 Multimodal information in live streaming

Given that the objective is to predict the ultimate sales of a live streaming, the current task can be conceptualized as a regression problem. Numerous factors can influence the final sales outcome. Thus, we try to collect five types of modal information provided by the live streaming platform, including text, image, numerical data, video, and audio. As mentioned above, the original data is categorized based on three entities, which are the anchor, the commodity and the live streaming room. It's noteworthy that we exclude audience data, as it's generated in real-time throughout the live streaming.

3.1.1 Anchor

The anchor plays a crucial mediating role in live streaming, serving not only as the creator of content but also as the bridge between the audience and the commodity. The anchor attracts a large number of audiences through personal characteristics, establishes an interactive relationship with them, and conveys information or entertainment content. Therefore, for the anchor entity, we primarily gather information reflecting their influence, as listed in Table 1.

Table. 1. Information from anchor entity

Attribute	Modality	Description
broadCasterField	Text	Field of an anchor in the platform
categoryName	Text	Category of an anchor in the platform
nick	Text	Nickname of an anchor
tags	Text	Tags of an anchor in the platform
HeadImage	Image	The avatar of an anchor
avg30DayView	Numerical	Number of times watched during live streaming in the past 30 days
avgComment	Numerical	average number of comments in all live streaming
avgLike	Numerical	average likes of comments in all live streaming
commentCount	Numerical	All comments count in all live streaming
fansCount	Numerical	The fan count of an anchor
followTopCount	Numerical	The fan count of those particularly focused on an anchor.
likeCount	Numerical	All likes count in all live streaming
score	Numerical	The ability score of an anchor in the platform

3.1.2 Commodity

The commodity entity mainly includes textual description, price, and image, as illustrated in Table 2. Textual description offers the various characteristics of the commodity, while price gives the commodity economic value and images provide a visual display. These attributes together constitute the information of the commodity and are important factors that affect the audience’s purchase decision.

Table. 2. Data from commodity entity

Attribute	Modality	Description
title	Text	Name of a commodity
itemRights	Text	Discount description of a commodity
brandName	Text	Brand of a commodity

shopName	Text	The shop that a commodity comes from
categoryName	Text	The category that a commodity belongs to
image	Image	The appearance image of a commodity
couponPrice	Numerical	Discounted price of a commodity
price	Numerical	Price of a commodity

3.1.3 Streaming room

The live streaming room serves as a core platform in the live streaming. And here, we mainly obtain its basic properties, including the name, tag, image and live streaming duration, as shown in Table 3. The name, tag, and image can provide the preliminary live content, which helps to attract audiences and stand out among many live streaming rooms. The duration records and presents the duration of the whole live streaming, which provides reference information for analyzing the workload of the anchor. Therefore, the attributes of streaming room also have an indispensable impact on the sales prediction in live streaming commerce.

Table. 3. Data from streaming room entity

Attribute	Modality	Description
title	Text	Name of a streaming room
tag	Text	Tags of a streaming room in the platform
headImage	Image	The avatar of a streaming room
duration	Numerical	Duration of a live streaming

3.1.4 Video-Audio information across entities and sales data

Video and audio information in live streaming simultaneously involves the data of three entities, i.e., the anchor, the commodity and the live streaming room. First, the appearance and voice of the anchor are captured during the streaming, translating into video and audio. At the same time, commodity information, such as image, description and price, is embedded into the video through floating windows. Additionally, details about the streaming room, like its background, are simultaneously recorded in the video. Thus, video and audio actually contain data from these three entities, providing the audience with richer information. Lastly, the sales data provides a direct reflection of the commodity’s sales performance and audiences purchasing behaviors during a live streaming. This data, crucially, is what we aim to predict, as detailed in Table 4.

Table. 4. Video-Audio information and sales data

Attribute	Modality	Description
Videos	Video	Compose a live video from multiple video clips
Audios	Audio	Compose a live audio from multiple audio clips
Sales	Numerical	Sales data, which is also the data that ultimately needs to be predicted

3.2 Feature extraction

In deep learning, feature representation is particularly critical. Proper feature representation enables deep neural network to more effectively capture and articulate the essence of data, thereby substantially enhancing the predictive performance and generalization capabilities. Therefore, for different modal information, we need appropriate methods to extract features.

Text content, such as the text description of the commodity and the live streaming room, provides valuable insights into the decision-making behavior of audiences. Considering the diversity of text, traditional text processing methods may be difficult to capture the rich semantics. In contrast, BERT [39] can interpret the context information in comments in this highly dynamic environment with its deep bidirectional architecture, and can be fine-tuned to adapt to text content under different entities. Therefore, we use BERT to extract text features in the live streaming context.

Images of the anchor, commodity, and live stream room are pivotal elements, which influence audience visual perception. RESNET [40] has successfully deepened the network structure and accurately captured complex visual information with its unique residual connection mechanism. More importantly, the features extracted by RESNET can better understand and deal with the image after training. Therefore, we use RESNET to extract image features.

Numerical data provides audiences with a more intuitive experience and plays a crucial role in sales forecasting. Because the multi-layer perceptron (MLP) has a deep structure and nonlinear activation function, it can efficiently learn and extract nonlinear features from complex data, which is a feed-forward neural network widely used in various tasks [41]. Therefore, using MLP to extract features helps reveal hidden relationships and potential insights in numerical data.

In the field of live streaming, video is not only the carrier of presenting content, but also the bridge of communication and interaction, which determines the quality of live streaming and audience experience. VideoMAE [42], as a state-of-the-art video feature extraction tool, can more accurately capture the details in video, thus better extracting the dynamic features. The accuracy of this feature extraction provides a deep and structured analysis foundation for live streaming content.

Serving as one of the core carriers of live streaming content, audio reflects the emotions, tone, and intent of the anchor. Traditional methods for audio feature extraction often rely on predefined rules and manually crafted features, which can lead to an inability to fully capture the deeper intricacies of the audio. In contrast, wav2vec [43], an unsupervised audio feature extraction model based on deep learning, can automatically capture higher-order statistical characteristics and obtain rich semantic representation of audio. Therefore, leveraging wav2vec for extracting audio features helps to analyze and understand the live content more accurately.

3.3 Feature fusion

Feature fusion has attracted increasing attention in multi-source data processing, especially in live streaming media. The interaction between different modal information can better serve the live streaming analysis. Therefore, in the feature fusion stage, we first divide all the information into three entities, i.e., anchor, commodity, and live streaming room. The Transformer architecture is used to handle the fusion and alignment of text, images, and numerical information under the same entity. Subsequently, the feature representations of these three entities are integrated with the video and audio features to attain a more accurate multimodal feature representation that impacts live sales. This process involves three distinct Transformer structures, which will be described in detail below.

3.3.1 Multimodal Transformer

The Multimodal Transformer is designed specifically for anchors, live streaming room, video and audio. First, for the two entities, anchor and live streaming room, we extracted text, numerical, and image features, denoted as $X_t, X_n, and\ X_i$, respectively. To ensure consistency in feature dimensions of different modalities, we introduce a fully connected layer for each modality, as shown in eq. (1) to eq. (3). The aim is to facilitate interaction among features from the three distinct modalities within a unified space. A target dimension $\widehat{D}$ is set, followed by transformation for each modality.

$$F_t = W_t \cdot X_t + b_t \tag{1}$$

$$F_n = W_n \cdot X_n + b_n \tag{2}$$

$$F_i = W_i \cdot X_i + b_i, \tag{3}$$

where $X_t \in R^{B\times Lt\times Dt}$, $X_n \in R^{B\times Ln\times Dn}$, $X_i \in R^{B\times Li\times Di}$, $W_t \in R^{Dt\times \widehat{D}}$, $W_n \in R^{Dn\times \widehat{D}}$, $W_i \in R^{Di\times \widehat{D}}$, $b_t, b_n, b_i \in R^{\widehat{D}}$, B denotes batch size, L represents the length of the modality, D indicates the dimensionality of the modality feature.

Furthermore, for different modalities pertaining to the same entity, we obtain the final input feature representation F by stacking these modal features together, as shown in eq. (4).

$$F = [F_t, F_n, F_i]^T, \tag{4}$$

where $F \in R^{B\times(Lt+Ln+Li)\times D}$.

Finally, we leverage the Transformer Encoder coupled with self-attention mechanisms to fuse the features from the three modalities. This architecture takes into consideration the interrelations and dependencies among features from different modalities, producing a more powerful and enriched unified feature representation. The module in the Transformer structure comprises a multi-head self-attention mechanism and a feed-forward neural network, each accompanied by a residual connection and layer normalization, as depicted in eq. (5) to eq. (8). Our Multimodal Transformer is composed of six such stacked modules, where the multi-head self-attention mechanism in each module has eight heads.

$$Self-Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

$$M=LayerNorm\left(F+Self-Attention\left(FW_q,FW_k,FW_v\right)\right) \tag{6}$$

$$FeedForward(Z)=ReLU(ZW_1+b_1)W_2+b_2 \tag{7}$$

$$Out=LayerNorm\big(M+FeedForward(M)\big) \tag{8}$$

For video and audio content, they are comprised of multiple segments that together form a complete live streaming. Due to the varying durations of different live streaming, the number of constituent segments differs. This variability results in an inconsistent number of features extracted from each live streaming, and these features are essentially unrelated. Therefore, it is necessary to further fuse them based on the Multimodal transformer structure to obtain complete video or audio features. Considering that the number of segments is inconsistent, padding is employed to achieve vectors of fixed length. This allows the sequence of a sample to be input into the Transformer as a matrix. Moreover, it's ensured that the padding portion doesn't contribute to the self-attention computation. Therefore, eq. (5) is modified to introduce mask mechanism to adapt the feature fusion of video and audio, as shown in eq. (9).

$$Self-Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\times mask\right)V \tag{9}$$

3.3.2 Multimodal QuadTransformer

For commodity, on the one hand, the modalities used to describe commodity information include text, numerical data, and images. On the other hand, the number of commodities sold in different live streaming varies. Thus, the final input for commodity information is represented as $F \in R^{B \times M \times (Lt+Ln+Li) \times D}$, where B stands for batch size, M represents the number of commodities, Lt indicates the count of text, Ln is the number of numerical data, Li denotes the count of images, and D signifies the feature dimension.

However, the conventional Transformer architecture only accepts three-dimensional tensor inputs and cannot handle four-dimensional tensors. Consequently, we propose the Multimodal QuadTransformer for commodity entity, facilitating the fusion of different modal features within commodities and the fusion between features of different commodities. Therefore, modifications are made to the Transformer structure, allowing it to first delve deeply into each modality of every commodity, and then consider the overall commodity relationship. This approach will yield a richer and more contextual representation of commodity entity.

a. The fusion of different modal features within a commodity

In live streaming scenario, for each commodity m ranging from 1 to M, i.e., the $\mathrm{m^{th}}$ slice of F. The slice is then passed through three fully connected layers, denoted as W_q, W_k and W_v, and generate of three distinct vectors, Q_m, K_m and V_m respectively. Subsequently, we leverage an attention mechanism to compute the fused representation of multi-modal features for the $\mathrm{m^{th}}$ commodity, as illustrated in eq. (10) to eq. (12).

$$Q_m,\ K_m,\ V_m = F_{:,m,:,:}W_q,\ F_{:,m,:,:}W_k,\ F_{:,m,:,:}W_v \tag{10}$$

$$Self - Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{11}$$

$$X_m = LayerNorm\left(F_{:,m,:,:} + Self - Attention(Q_m, K_m, V_m)\right) \tag{12}$$

b. The fusion of features among commodities

For the final feature representation X_m of each commodity m, we obtain an integrated commodity feature $\hat{X}_m$ by averaging across all modalities. The aim is to amalgamate information from each modality into a unified representation, thereby offering a consistent view for subsequent processes. To comprehend the relationship between each commodity and others, we employ a self-attention mechanism. This implies that the representation of each commodity is updated based on the representations of other commodities. In the live streaming context, this may assist in discerning which commodities frequently co-occur or which commodities are similar in content or price. The representation of each commodity is further refined through a feed-forward neural network to extract more semantic features. To enhance the model's training stability and performance, residual connections and layer normalization are added after each step, as shown in eq. (13) to eq. (18).

$$\hat{X}_m = \frac{1}{Lt + Ln + Li} \sum X_m \tag{13}$$

$$X = Concat\left(\hat{X}_1, \hat{X}_2, \dots, \hat{X}_M\right) \tag{14}$$

$$Self - Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} \times mask\right)V \tag{15}$$

$$M = LayerNorm\left(X + Self - Attention\left(XW_q, XW_k, XW_v\right)\right) \tag{16}$$

$$FeedForward(Z) = ReLU(ZW_1 + b_1)W_2 + b_2 \tag{17}$$

$$Out = LayerNorm\left(M + FeedForward(M)\right) \tag{18}$$

3.3.3 Multi-Entity Transformer

The ultimate goal of the Multi-Entity Transformer is to fuse the output results of the Multimodal Transformer and the Multimodal QuadTransformer at the entity level, aiming to obtain the final feature representation of live streaming. As previously mentioned, the three entities are described by three modality features, i.e., text, numerical data, and image, and video and audio cover the information of three entities. However, these features are different to some extent, so it is necessary to further fuse these features to get the final feature representation of live streaming. Here we still use the transformer structure to extract features on different entities, and name it Multi-Entity Transformer.

More specifically, live streaming platform provide audiences with information about anchor, commodity, and live streaming room. All the information is complementary and affects audiences' behavior together.

Therefore, through the Multi-Entity Transformer, we can capture the interactions and relationships between these entities, leading to a more accurate and comprehensive representation of live-streaming features.

3.4 Sale prediction

Our objective is to explore the nonlinear mapping mechanism from integrated multi-modal features to sales. Many studies have shown that deep neural network (DNN) has excellent performance in many fields. Based on this, we choose DNN as the core predictor, anticipating accurate prediction outputs. The input received by the predictor is the fusion feature of multimodal information, which gives it more rich and comprehensive data insight. It is worth noting that after two fusions, various modal features have been effectively mapped into a unified semantic space. Therefore, through the complex structure of DNN, we can deeply explore the potential patterns and their internal associations in the fusion features, so as to achieve more accurate prediction. During the construction of DNN, we select RELU as the main activation function, and employ dropout technology to prevent over fitting.

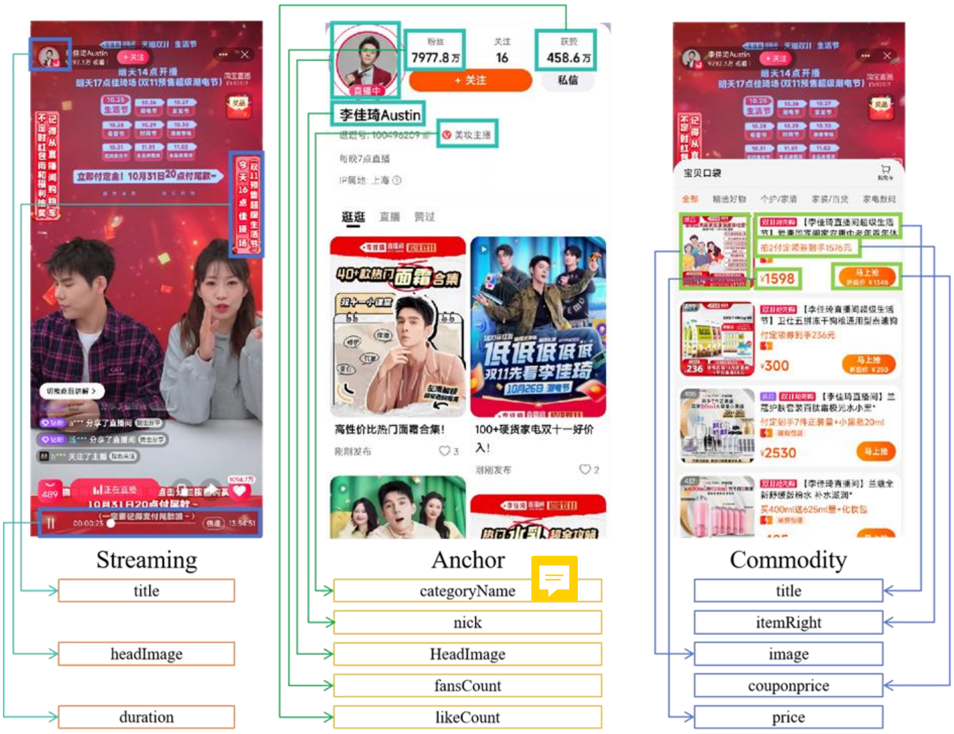

Fig.2. Example of live streaming

4. Experiments

4.1 Data description

This study utilizes a real-world dataset sourced from Taobao Live. As a leading live-stream shopping platform in China, Taobao Live commands a substantial share of the apparel sales in the live-streaming commercial market. Figure 2 presents a sample of the live streaming data, showcasing partial information related to anchor, commodity and live streaming room.

The data collection period spans from October 1, 2021, to October 31, 2021. Notably, certain key data, such as sales volume, is not directly acquired from Taobao Live but is sourced from a third-party data provider named *huntun*. Among the 15326 preliminarily collected live streaming data points, we exclude the data that lacks information from any of the modalities in order to ensure accurate modality comparisons. Moreover, considering that flagship stores and official stores do not have designated anchors, their relevant data are also excluded. Consequently, the experiment is conducted based on 920 live streaming data points. The dataset is divided into training dataset and test dataset, accounting for 80% and 20% respectively. The statistics of the dataset are listed in Table 5.

Table. 5. Statistics of the datasets.

dataset	size	log(sales)				
		Max	Min	Mean	Median	Standard Deviation
Training set	736	7.7366	1.3075	4.9927	4.8483	1.0733
Test set	184	8.0105	1.6981	5.0923	5.0029	1.0846

In this study, we comprehensively consider the feature extraction of multimodal information, including text, numerical data, image, video, and audio. For the textual component, we initially carry out a series of preprocessing steps that include tokenization, removal of stop words, and handling special characters. Subsequently, using the BERT pre-trained model maps each piece of text to a fixed-length semantic vector. In terms of image processing, each image undergoes uniform size adjustment and normalization. The ResNet50 pre-trained model is then employed to capture its deep features. For video, we choose the

VideoMAE pre-trained model to precisely extract its temporal and spatial characteristics, and provide rich feature description for subsequent analysis tasks. Regarding audio, we standardize its sampling rate to cater to the Wav2Vec model's requirements and utilize this model to extract high-dimensional features directly from the raw waveforms. For sales and other numerical data, we apply logarithmic transformation for standardization.

4.2 Experimental settings

All the work is carried out in the python environment, and all the experiments involving deep learning are based on PyTorch. The hyperparameters considered include the learning rate, dropout rate, embedding dimension of the Transformer architecture, and the number of multi-head attention mechanisms. Given that the feature extraction phase is based on pre-trained models, an excessively high learning rate could lead to model divergence. Therefore, the examined range for the learning rate is {5e-7, 1e-6, 5e-6, 1e-5, 5e-5}. To assess the impact of dropout rate on results, we examine a range of {0.1, 0.2, 0.3, 0.4, 0.5}. For the Transformer architecture, the range for embedding dimensions is {256, 512, 768}, and the range for the number of multi-head attention mechanisms is {4, 8, 16}. The finalized configurations can be found in Table 6. We employe the Adam optimizer with a learning rate decay, which has decay step size of 50 and a multiplicative decay factor gamma of 0.98.

Table. 6. Hyperparameters settings

Hyperparameters	Range of search space	Final Setting
Learning rate	{5e-7, 1e-6, 5e-6, 1e-5, 5e-5}	1e-6
Dropout	{0.1, 0.2, 0.3, 0.4, 0.5}	0.1
Embedding dimension	{256, 512, 768}	512
Heads	{4, 8, 16}	8

To better evaluate our model, we select several advanced multimodal fusion techniques as baseline models, including **MFB** [32], **TFN** [33], **MFH** [34], **VTFSA** [35] and **CMMT** [36]. **MFB (Multimodal Factorized Bilinear Pooling)** is a prominent method in the field of multimodal fusion. The essence of

MFB is that it employs a bilinear pooling technique to fuse features from different modalities. **MFH (Multimodal Factorized High-order Pooling)** is characterized by implementing high-order interactive fusion through stacking multiple bilinear pooling operations. **TFN (Tensor Fusion Networks)** uses tensor decomposition strategizes to model interactions between each modality and hidden representations. **VTFSA** primarily employs a self-attention mechanism to efficiently fuse features from different modalities. Here, we follow the original authors' design philosophy and adapt the model structure to ensure its optimal application for our sales prediction task in live streaming commerce. **CMMT** utilizes a cross-attention strategy to effectively integrate different modalities and reveal complex relationships of them. Notably, as **CMMT** is originally designed for multimodal sentiment analysis tasks, we make some corresponding adjustments to better suit our research needs.

The main difference between these baseline models and our model lies in the multimodal fusion part, while the feature extraction, DNN predictor and other parts are completely consistent for fair comparison.

4.3 Evaluation metrics

Since our goal can be regarded as a regression task, this study uses the following evaluation metrics to measure the comprehensive performance of our model. **MAE** represents the average of the absolute discrepancies between actual and predicted values. **MSE** denotes the mean of squared differences between the actual and predicted values. **MAPE** expresses the average absolute difference between actual and predicted values as a percentage of the actual values. **R-squared Score (r2_score)** is the coefficient of determination, reflecting the degree of fitting of the model. **Explained Variance Score (EVS)** is a measurement used to quantify the explanatory power of a model for the variance of the dependent variable. These metrics together provide us with a comprehensive understanding and help us evaluate the accuracy and robustness of the model.

4.4 Experimental results

4.4.1 Performance of MEMF

Here we show the changes in the loss function during the training process of the model proposed by us and the evaluation of five metrics on the test dataset, as shown in Fig. 3. As can be observed from the figure, the training loss rapidly decreases in the initial stage, which means that the model can effectively capture key patterns and structures from the data in the early stages of learning. However, as the training period increases, the decrease in training loss gradually slows down and stabilizes in the later stages. The Test MSE, Test MAE, and Test MAPE provide us with information about the performance of the model on test dataset. These three metrics also showed a rapid downward trend during the initial stages of training, validating the significant improvement in the model's generalization ability during the early stages. Similar to the training loss, these three metrics also showed a trend towards stability during the later stages. R2_score and EVS provide information about the quality of model fitting, and their values increase continuously as the training period progresses until they reach a balance. The final results demonstrate that the model has good fitting effects on multimodal data in the live streaming commerce.

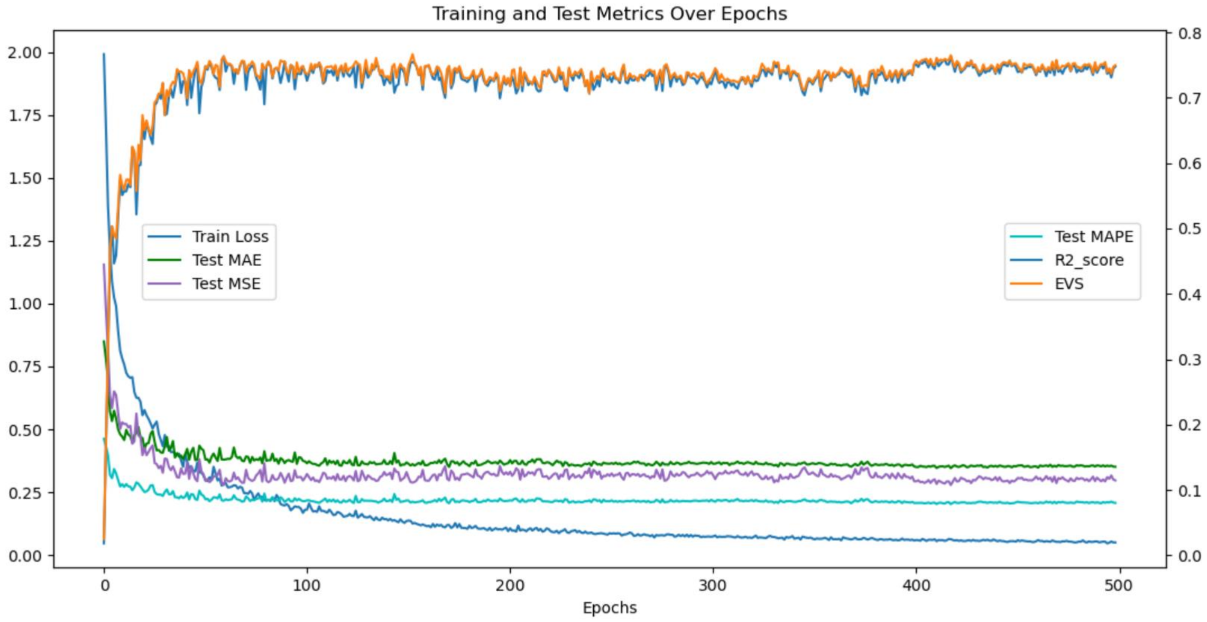

Fig.3. Training and Test Metrics Over Epochs

Additionally, we also show the scatter plot density distribution of the model's predicted values and actual observed values, as shown in Fig. 4. Through the 1:1 line, we can intuitively observe that the predicted values are close to the true values. The data points are mainly concentrated in the darker regions, which means that the model's predictions are more accurate in these regions. However, there are also some data points that are far away from the 1:1 line, representing the model’s large prediction error in some cases. Further, through evaluation metrics, we can quantify the model’s prediction ability. The R2_score value of 0.7653 indicates that the model explains 76.53% of the data variability and performs relatively well.

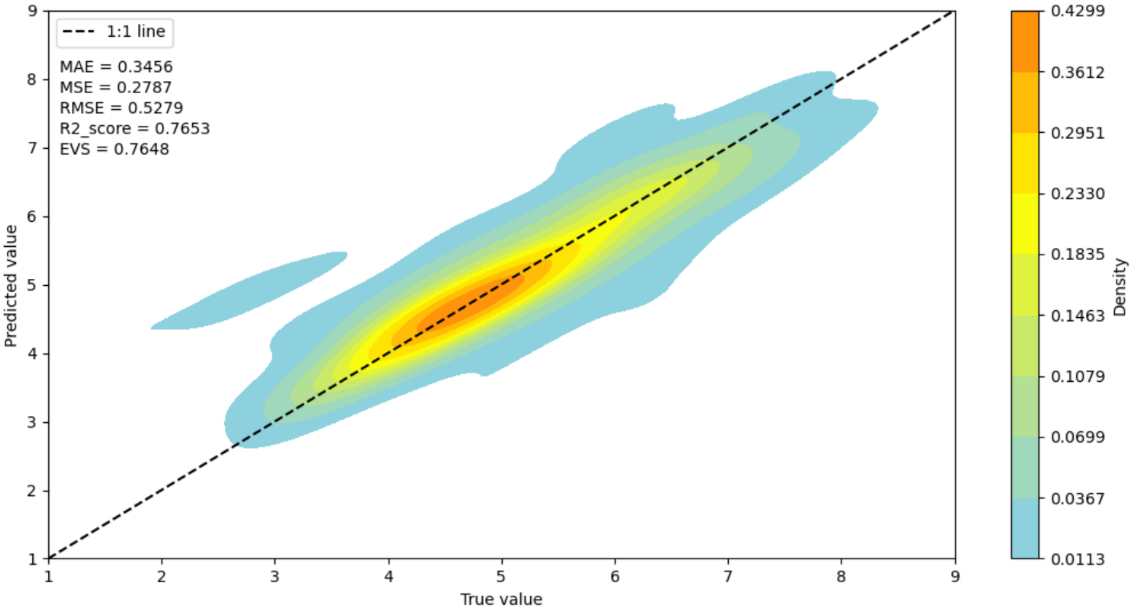

Fig.4. Scatter density plot between the true and predicted values of regression prediction results

4.4.2 Effectiveness evaluation of Muti-Entity Transformer

In our model, we use a Multi-Entity Transformer at the entity level to fuse and obtain the final feature representation. To verify the effectiveness of this method, we design a comparative experiment for different fusion methods at the entity level. Since Multimodal Transformer and Multimodal QuadTransformer represent the multi-modal features of the anchor, commodity, and live streaming room, as well as the video and audio features, in 768 dimensions, we select four common methods for fusion at the entity level for comparison.

The **Mean** method fuses features of different entities by averaging, without requiring any additional parameters, which reduces the complexity of the model.

The Fully Connected layer (FC) can learn any relationship between features between entities, providing greater flexibility for fusion.

LSTM can learn and retain more complex information patterns through its gate mechanism, and can adapt to different lengths of input to generate fixed-size output representations.

CNN can capture local features and patterns of entities through different sizes of convolution kernels. By stacking multiple convolution layers, it can learn higher-level representations from raw features.

Table. 7. Results of different fusion methods on entity level

Method	MAE	MSE	MAPE	R2_score	EVS
Mean	0.4105	0.3425	0.0925	0.7089	0.7139
FC	0.4050	0.3505	0.0916	0.7021	0.7097
LSTM	0.4018	0.3011	0.0896	0.7440	0.7479
CNN	0.3847	0.2996	0.0857	0.7453	0.7474
Muti-Entity Transformer	**0.3456**	**0.2787**	**0.0786**	**0.7631**	**0.7648**

From Table 7, we can observe that, due to its simple structure, the Mean model cannot effectively fuse the features of the entity layer. However, because of parameter training, FC, LSTM, and CNN can better capture and interact with the features of different entity layers. Notably, the Multi-Entity Transformer outperforms all in all evaluation metrics. This validates that our decision to use the Muti-Entity Transformer for entity level fusion is reasonable and effective.

4.4.3 Performance comparison among different sequence of feature fusion

Considering that MEMF first fuses different modalities under different entities, and then fuses them at the entity level. To verify the effectiveness of this sequence, we design another experiment, where the same modalities are fused according to entities first, and then the different modalities are fused. The results are shown in Fig. 5, and it’s evident that the fusion sequence adopted by MEMF is more rational. MEMF exhibits lower error values in MAE, MSE, and MAPE metrics, while it also demonstrates higher scores in the R2_score and EVS metrics. This strongly demonstrates the rationale and superiority of this sequence.

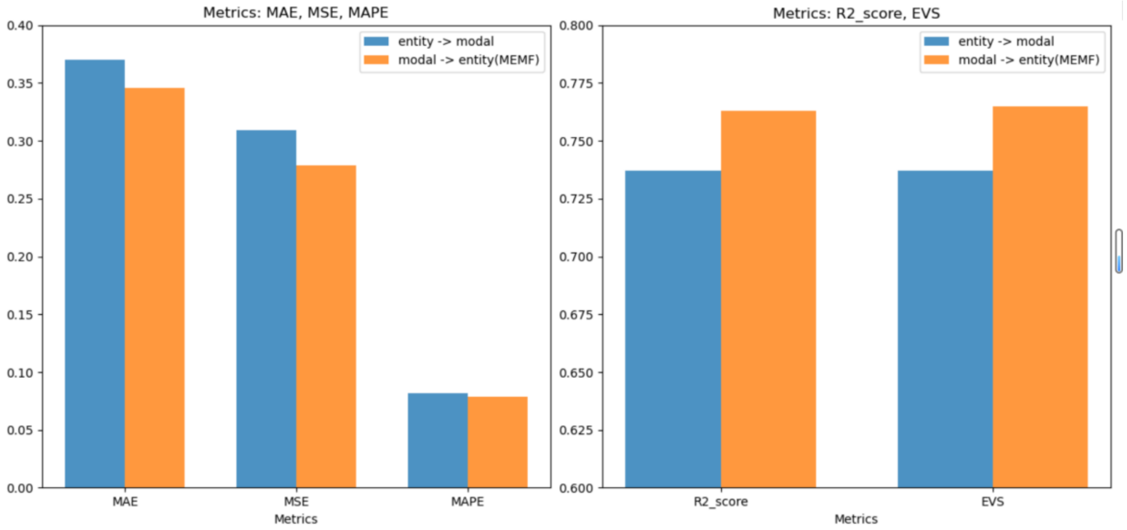

Fig. 5. Comparison among different sequence of feature fusion

4.4.4 Performance comparison on different entities feature

In the process of studying the live streaming sales prediction, we recognize that different entity-level features may have varying degrees of impact on prediction results. To clarify the importance of entity-level features, we conduct a series of experiments. Entity-level features include anchor, commodity, live room, videos, and audios. Since videos and audios cover both anchor, commodity, and live streaming room, here we separately analyze videos and audios.

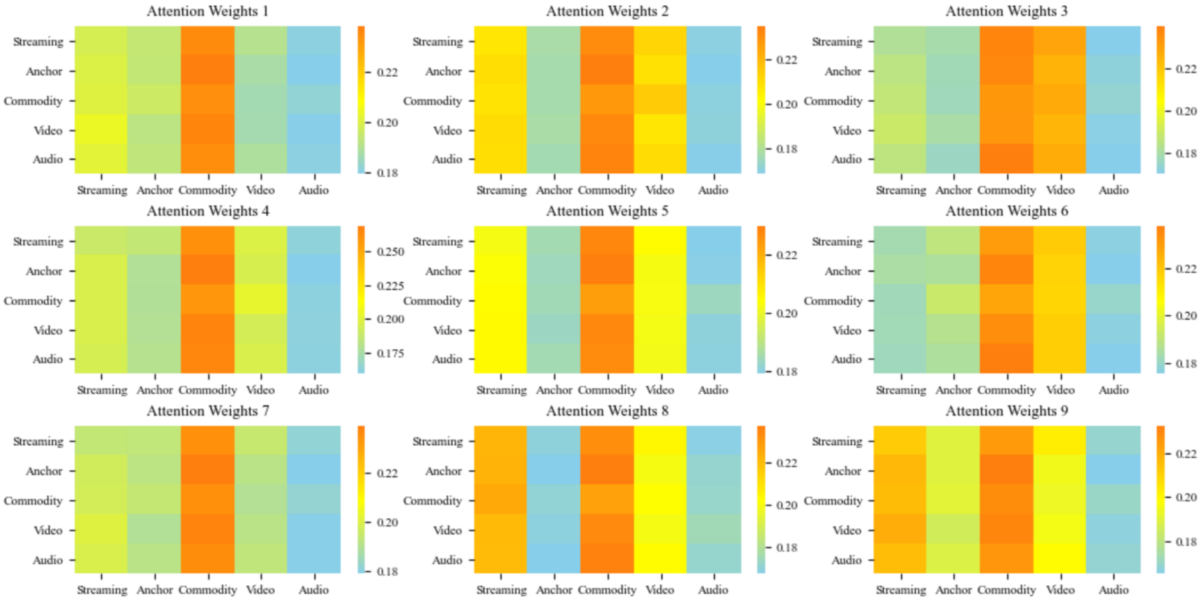

Fig. 6. Heatmap of the attention weight matrix

Firstly, given that the attention mechanism in the Transformer model allows the model to assign different weights to inputs at different positions when processing sequence data. In order to better understand how the model focuses on different parts of the input, we randomly obtain 9 live streaming data points from the test dataset and use the visualization method of heat maps to reveal the internal weights of the Multi-Entity Transformer on the entity level, as shown in Fig. 6. It can be seen that for different data points, the influence of different entities varies. But overall, the information of commodity entity has a significant impact on the model's performance, while the information of audio has a smaller impact.

Furthermore, we explore the overall impact of different entities on the final sales prediction in live streaming. To more accurately measure the impact of these entities, we conduct a series of experiments, including experiments that exclude specific entities from the complete model one by one, and compare their prediction results with the complete model. The results are shown in Table 8.

Table. 8. Impact of different entities

Description	MAE	MSE	MAPE	R2_score	EVS
All without Streaming	0.3950	0.3855	0.0930	0.6723	0.6780
All without Anchor	0.4007	0.4077	0.0956	0.6534	0.6833
All without Commodity	0.4308	0.3726	0.0956	0.6832	0.6895
All without Audio	0.3622	0.3038	0.0823	0.7418	0.7514
All without Video	0.4199	0.3925	0.0946	0.6664	0.6857
All	**0.3456**	**0.2787**	**0.0786**	**0.7631**	**0.7648**

From Table 8, we can clearly observe that there are indeed differences in the impact of anchor, commodity, and the live streaming room on sales prediction. Among them, using all the entity data for prediction has the best performance on all evaluation metrics, indicating that when integrating all entities for sales prediction, the model's prediction effect is optimal. When comparing the differences between each entity, we find that commodity information is the main factor affecting the accuracy of the prediction. Although the information of the anchor and the live streaming room also has an impact on the prediction results, the influence is relatively small, corresponding to Figure 5. This also confirms that in live streaming, commodity information is important. Factors such as commodity quality, price, and brand are all key factors in determining commodity sales.

4.4.5 Performance comparison of the methods

To verify the effectiveness of the proposed model, we compare it with several popular multimodal fusion methods, as shown in Table 9.

Table. 9. Results of different fusion methods

Method	MAE	MSE	MAPE	R2_score	EVS
MFB	0.3791	0.3199	0.0827	0.7281	0.7306
MHB	0.4010	0.3389	0.0865	0.7119	0.7120
TFN	0.4341	0.3978	0.0964	0.6619	0.6619
CMMT	0.3845	0.3104	0.0833	0.7362	0.7375
VTFSA	0.3577	0.3083	0.0797	0.7379	0.7516
MEMF	**0.3456**	**0.2787**	**0.0786**	**0.7631**	**0.7648**

As seen in Table 9, compared to other baseline models, our proposed model demonstrates superior performance in terms of MAE, MSE, MAPE, R2_score, and EVS evaluation metrics. In comparison to the second-best model, MEMF reduces errors by 3.4%, 9.7%, and 1.4% in the MAE, MSE, and MAPE metrics respectively, while also improves the fitting degree by 3.4% and 1.8% in R2_score and EVS metrics respectively. These results verify the effectiveness of our model design and implementation approach.

The main reason is that our model performs two multimodal fusions. First, under the three entities, we use the Transformer structure to process three different modal information under the same entity, which enables the model to effectively capture and fuse different modalities of relationships. Second, we use the Transformer model to integrate the three entities of anchor, commodity, and live room, as well as the video and audio features, enabling the model to more comprehensively capture the full-modal feature representation of the live streaming. It also verifies that incorporating the same modality information into the model uniformly results in the loss of some information, which affects the final prediction performance.

4.4.6 Performance comparison of different feature extraction methods

Considering that different modal feature extraction methods may lead to changes in the performance, we adopt more feature extraction methods for image and text modalities to evaluate MEMF and other models. Specifically, we select current popular pre-trained models for feature extraction of images and texts. The pre-trained models for images include Resnet50, DenseNet121, Resnet18, and Resnet101 and the pre-trained models for text include BERT and XLNET. As shown in Table 10, compared with the current popular multimodal fusion methods (CMMT and VTFSA), MEMF has the best performance under different feature extraction methods, indicating that our model is robust.

Table. 10 The prediction performance of different feature extraction methods

	Pre-train models	Model	MAE	MSE	MAPE	R2_score	EVS
Image	Resnet50	CMMT	0.3845	0.3104	0.0833	0.7362	0.7375
		VTFSA	0.3577	0.3083	0.0797	0.7379	0.7516
		MEMF	**0.3456**	**0.2787**	**0.0786**	**0.7631**	**0.7648**
	DenseNet121	CMMT	0.3753	0.3053	0.0854	0.7405	0.7473
		VTFSA	0.3658	0.3206	0.0805	0.7274	0.7280
		MEMF	**0.3642**	**0.2621**	**0.0792**	**0.7772**	**0.7808**
	Resnet18	CMMT	0.3775	0.2824	0.0825	0.7599	0.7609
		VTFSA	0.37811	0.2978	**0.0795**	0.7469	0.7518
		MEMF	**0.3578**	**0.2767**	0.0817	**0.7648**	**0.7696**
	Resnet101	CMMT	0.3933	0.2908	0.0851	0.7528	0.7546
		VTFSA	0.3922	0.3258	**0.0837**	0.7231	0.7257
		MEMF	**0.3900**	**0.2855**	0.0880	**0.7573**	**0.76544**
Text	BERT	CMMT	0.3845	0.3104	0.0833	0.7362	0.7375
		VTFSA	0.3577	0.3083	0.0797	0.7379	0.7516
		MEMF	**0.3456**	**0.2787**	**0.0786**	**0.7631**	**0.7648**
	XLNET	CMMT	0.4004	0.3238	0.0873	0.7248	0.7250
		VTFSA	0.3757	0.3376	0.0817	0.7130	0..7131
		MEMF	**0.3458**	**0.2946**	**0.0802**	**0.7496**	**0.7510**

4.5 Discussion

In this study, we have proposed the Multi-Entity Multimodal Fusion framework (MEMF) to enhance sales prediction in live streaming commerce. Our framework is innovative in methodology, introducing a novel entity-level fusion perspective on the basis of multimodal fusion, which more effectively captures the complexity of live streaming than previous models.

Methodologically, our framework offers a unique contribution to the decision support systems literature by addressing the challenges of fusing more fusing more modal information from different entities. Unlike existing multimodal approaches that may ignore the differences in modal information between entities, the MEMF framework acknowledges and leverages this heterogeneity. This approach has resulted in improved predictive accuracy, as detailed in Table 9 and Table 10, where our model exhibits better performance compared to existing multimodal fusion methods. The innovations of our framework lie in its ability to synthesize data from various modalities, as well as the entity-level fusion, which is carried out specifically for modalities under different entities. These have provided a more nuanced and comprehensive feature representation, leading to the increase in predictive precision.

The managerial implications of this study are manifold. Firstly, for practitioners in live streaming commerce, our framework suggests that strategic emphasis should be placed on the selection of commodities, given their substantial impact on consumer purchase intentions. Merchants and streamers can utilize our MEMF framework to identify the most popular commodities for the audience, thereby informing inventory selection and promotional tactics. Secondly, our framework emphasizes the importance of the role of anchors. Anchors typically serve as a bridge between commodities and consumers, playing a role beyond displaying commodities, and their personal abilities can affect consumers' purchasing decisions. Our framework can provide a platform for them to optimize their performance in a targeted manner, such as improving presentation techniques. Thirdly, the impact of streaming room is minimal, indicating that while the aesthetics of the environment may contribute to audiences' enjoyment, they are less influential on purchasing decisions. This is a valuable insight for merchants and streamers, suggesting that investments in the streaming environment should be balanced with other more important factors.

5. Conclusions

This paper has proposed a Multi-Entity Multimodal Fusion Framework to predict sales in live streaming commerce. We extract text, image, numerical, video, and audio features from live streaming, and divide them into three entities, i.e., anchor, commodity, and live room. Through the fusion of multimodal information under multiple entities, we achieve accurate sales prediction. As an important component of MEMF, Multimodal Transformer can better fuse multimodal information under different entities. Multimodal QuadTransformer aims to better obtain different modality features within commodities and feature relationships between commodities, providing richer and more contextual commodity entity representations. And Mul-Entity Transformer is a comprehensive representation of multiple entities, to obtain the final feature representation of live streaming. Empirical analysis shows that our proposed MEMF outperforms other existing multimodal fusion methods. MEMF not only provides more accurate and stable prediction results, but also can handle more complex and multimodal data. Its efficient model structure and flexible multimodal fusion strategy make it perform well in live sales prediction.

Of course, there are some limitations in this study. First, although we use multimodal data of video and audio, we still adopt a macro perspective in feature extraction, ignoring details such as the host's gestures and facial expressions. Second, we measure the influence of different entities through visualizing attention weights and experiments, but there is still a lack of deep explanation on how these entities specifically affect sales prediction. Third, our model is mainly designed for the live streaming commerce, but its effectiveness and adaptability in other scenarios remains unknown. Therefore, future research directions will focus on further improving the model, deeply mining information features, and exploring the potential of MEMF in more widely applicable scenarios.

References

[1] Zhang, M., Liu, Y., Wang, Y., & Zhao, L. (2022). How to retain customers: Understanding the role of trust in live streaming commerce with a socio-technical perspective. Computers in Human Behavior, 127, 107052.

[2] Shin, H., Oh, C., Kim, N. Y., Choi, H., Kim, B., & Ji, Y. G. (2023). Evaluating and eliciting design requirements for an improved user experience in live-streaming commerce interfaces. Computers in Human Behavior, 107990.

[3] Hu, M., & Chaudhry, S. S. (2020). Enhancing consumer engagement in e-commerce live streaming via relational bonds. Internet Research, 30(3), 1019-1041.

[4] Xu, W., Zhang, X., Chen, R., & Yang, Z. (2023). How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce. Decision Support Systems, 113984.

[5] Xin, B., Hao, Y., & Xie, L. (2023). Strategic product showcasing mode of E-commerce live streaming. Journal of Retailing and Consumer Services, 73, 103360.

[6] Peng, J., Zhang, J., & Nie, T. (2023). Social influence and channel competition in the live-streaming market. Annals of Operations Research, 1-29.

[7] Gustriansyah, R., Ermatita, E., & Rini, D. P. (2022). An approach for sales forecasting. Expert Systems with Applications, 207, 118043.

[8] Xi, C., Lu, G., & Yan, J. (2020). Multimodal sentiment analysis based on multi-head attention mechanism. In Proceedings of the 4th international conference on machine learning and soft computing (pp. 34-39).

[9] Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., ... & Huang, Y. (2022). Video sentiment analysis

with bimodal information-augmented multi-head attention. Knowledge-Based Systems, 235, 107676.

[10] 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.

[11] 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.

[12] Chen, X., Shen, J., & Wei, S. (2023). What reduces product uncertainty in live streaming e-commerce? From a signal consistency perspective. Journal of Retailing and Consumer Services, 74, 103441.

[13] Lo, P. S., Dwivedi, Y. K., Tan, G. W. H., Ooi, K. B., Aw, E. C. X., & Metri, B. (2022). Why do consumers buy impulsively during live streaming? A deep learning-based dual-stage SEM-ANN analysis. Journal of Business Research, 147, 325-337.

[14] Lu, B., & Chen, Z. (2021). Live streaming commerce and consumers' purchase intention: An uncertainty reduction perspective. Information & Management, 58(7), 103509.

[15] Chen, H., Dou, Y., & Xiao, Y. (2023). Understanding the role of live streamers in live-streaming e-commerce. Electronic Commerce Research and Applications, 59, 101266.

[16] Mikalef, P., Sharma, K., Chatterjee, S., Chaudhuri, R., Parida, V., & Gupta, S. (2023). All eyes on me: Predicting consumer intentions on social commerce platforms using eye-tracking data and ensemble learning. Decision Support Systems, 114039.

[17] Khoi, N. H., Le, A. N. H., & Dong, P. N. (2023). A moderating–mediating model of the urge to buy impulsively in social commerce live-streaming. Electronic Commerce Research and Applications, 60, 101286.

[18] Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average

and quantile regression for daily food sales forecasting. International Journal of Production Economics, 170, 321-335.

[19] Taylor, J. W. (2011). Multi-item sales forecasting with total and split exponential smoothing. Journal of the Operational Research Society, 62(3), 555-563.

[20] Joseph, R. V., Mohanty, A., Tyagi, S., Mishra, S., Satapathy, S. K., & Mohanty, S. N. (2022). A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting. Computers and Electrical Engineering, 103, 108358.

[21] Pan, S. Y., Liao, Q., & Liang, Y. T. (2022). Multivariable sales prediction for filling stations via GA improved BiLSTM. Petroleum Science, 19(5), 2483-2496.

[22] Vallés-Pérez, I., Soria-Olivas, E., Martínez-Sober, M., Serrano-López, A. J., Gómez-Sanchís, J., & Mateo, F. (2022). Approaching sales forecasting using recurrent neural networks and transformers. Expert Systems with Applications, 201, 116993.

[23] Bogaert, M., Ballings, M., Van den Poel, D., & Oztekin, A. (2021). Box office sales and social media: A cross-platform comparison of predictive ability and mechanisms. Decision Support Systems, 147, 113517.

[24] Khatiwada, A., Kadariya, P., Agrahari, S., & Dhakal, R. (2019). Big Data Analytics and Deep Learning Based Sentiment Analysis System for Sales Prediction. In 2019 IEEE Pune Section International Conference (PuneCon) (pp. 1-6). IEEE.

[25] Tu, Z., Yang, W., Wang, K., Hussain, A., Luo, B., & Li, C. (2023). Multimodal salient object detection via adversarial learning with collaborative generator. Engineering Applications of Artificial Intelligence, 119, 105707.

[26] Cheung, T. H., & Lam, K. M. (2022). Crossmodal bipolar attention for multimodal classification

on social media. Neurocomputing, 514, 1-12.

[27] Rafailidis, D., Manolopoulou, S., & Daras, P. (2013). A unified framework for multimodal retrieval. Pattern Recognition, 46(12), 3358-3370.

[28] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), 423-443.

[29] Zhu, S., Fang, Z., Wang, Y., Yu, J., & Du, J. (2019). Multimodal activity recognition with local block CNN and attention-based spatial weighted CNN. Journal of Visual Communication and Image Representation, 60, 38-43.

[30] Qin, W., Tang, J., Lu, C., & Lao, S. (2022). A typhoon trajectory prediction model based on multimodal and multitask learning. Applied Soft Computing, 122, 108804.

[31] Cai, Y., Zhang, Z., Ghamisi, P., Rasti, B., Liu, X., & Cai, Z. (2023). Transformer-based contrastive prototypical clustering for multimodal remote sensing data. Information Sciences, 649, 119655.

[32] Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 1821-1830).

[33] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

[34] Yu, Z., Yu, J., Xiang, C., Fan, J., & Tao, D. (2018). Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems, 29(12), 5947-5959.

[35] Cui, S., Wang, R., Wei, J., Hu, J., & Wang, S. (2020). Self-attention based visual-tactile fusion

learning for predicting grasp outcomes. IEEE Robotics and Automation Letters, 5(4), 5827-5834.

[36] Yang, L., Na, J. C., & Yu, J. (2022). Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. Information Processing & Management, 59(5), 103038.

[37] Gao, G., Liu, H., & Zhao, K. (2023). Live streaming recommendations based on dynamic representation learning. Decision Support Systems, 169, 113957.

[38] Tian, R., Yin, R., & Gan, F. (2022). Exploring public attitudes toward live-streaming fitness in China: A sentiment and content analysis of China's social media Weibo. Frontiers in Public Health, 10, 1027694.

[39] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

[40] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[41] Rynkiewicz, J. (2019). Asymptotic statistics for multilayer perceptron with ReLU hidden units. Neurocomputing, 342, 16-23

[42] Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems, 35, 10078-10093.

[43] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.