Design and Research of Cross-Border E-Commerce Short Video Recommendation System Based on Multi-Modal Fusion Transformer Model

Yiran Hu*
School of Finance and Trade Management, Chengdu Industry and Trade College, Chengdu 611731, China

***Abstract*—This study designed a cross-border e-commerce short video recommendation system based on Transformer's multimodal analysis model. When mining associations, the model not only focuses on the relationships between modalities, but also improves semantic context by addressing contextual correlations within and between modalities. At the same time, the model uses a cross modal multi head attention mechanism for multi-level association mining, and constructs an association network interwoven with latitude and longitude. In the process of exploring the essential correlation between patterns and subjective emotional fluctuations, the potential context between patterns has been realized. Fully explore correlations and then more accurately identify the truth contained in the original data. In addition, this study proposes a self supervised single modal label generation method. When multimodal labels are known, it does not require complex deep networks and only relies on the mapping relationship between multimodal representations and labels to generate a single modal label. Modal labeling can achieve phased automatic labeling of single modal labels, and quantify the mapping relationship between modal representations and labels from the representation space to generate weak single modal labels. The study also achieved multimodal collaborative learning in the context of limited differential information acquisition due to incomplete labeling, fully utilizing multimodal information. The experimental results on classic datasets in the field of multimodal analysis show that it outperforms the baseline model in terms of accuracy and F1 score, reaching 98.76% and 97.89%, respectively.**

***Keywords*—Multimodal fusion; transformer model; cross-border e-commerce; short video recommendation system**

I. Introduction

With the advent of the era of big data, new social media such as DouYin, Weibo and YouTube will update a large amount of data content every day, in which there are not only objective descriptions of a certain thing, but also a large number of subjective expressions [1, 2]. Mining and identifying the information contained in these data can not only provide information assistance for big data forecasting applications such as financial market trend forecasting, product marketing status forecasting, and even US political election forecasting, but also provide information decision-making such as network public opinion analysis and digital social governance [3]. Providing technical support has extremely important application value and practical significance.

According to the existing research situation, traditional text analysis only uses words, phrases and their semantic associations to judge, which is not enough to identify complex information. Multimodal analysis adds acoustic and visual information on the basis of text information, and with the help of the association between multimodal data, it can show the information that may be hidden in text data, so as to achieve more accurate recognition [4, 5]. Taking ironic emotion recognition as an example, by extracting acoustic and visual information from human intonation and body movements, ironic information can be accurately recognized. Multimodal analysis has achieved remarkable results in dealing with understanding in various scenarios, and has attracted more and more researchers' attention [6, 7]. However, there are still some challenges in the research of multimodal analysis-multimodal association mining and multimodal collaborative learning.

In response to these challenges, this paper considers the association information between modes and contexts in the process of multimodal analysis based on deep learning technology. It uses the improved Transformer framework to mine intertwined and intricate associations to achieve tight coupling of multimodal data. Focusing on the mapping relationship between multimodal representations and sample labels, multimodal collaborative learning under unbalanced information distribution is realized with the help of a multi-task learning framework. Multimodal fusion is properly sorting and tightly coupling data from two or more modes. The most significant difference between multimodal analysis and traditional single-modal analysis is that the former can obtain more reliable prediction results with the help of information gained by multi-source data. According to the different stages, the existing multimodal fusion methods can be divided into three categories: feature-level, decision-level, and hybrid-level. Multimodal analysis based on Transformer and multi-task learning has essential application significance. However, its research results can also provide a basis and support for cross-media perceptual computing, analytical reasoning, and multimodal deep learning research in artificial intelligence. Has important research significance.

II. Multimodal Analysis based on Transformer

A. Transformer for Linguistics Guidance

The traditional multi-head attention mechanism is mostly applied to machine translation problems. When calculating the attention score, the operation can be performed parallelly to accelerate the training of the model [8, 9]. This paper applies this idea to the multi-modal problem, hoping to find the mapping relationship between multiple modes. Specifically, when using the attention mechanism to learn a mode, the text mode is used

*Corresponding Author.

as a guide to mine the association between various modes, and finally a linguistic-guided Transformer (LGT) is constructed. LGT includes Multi-Head Attention (MHA) and Forward Neural Networks (FNN) [10].

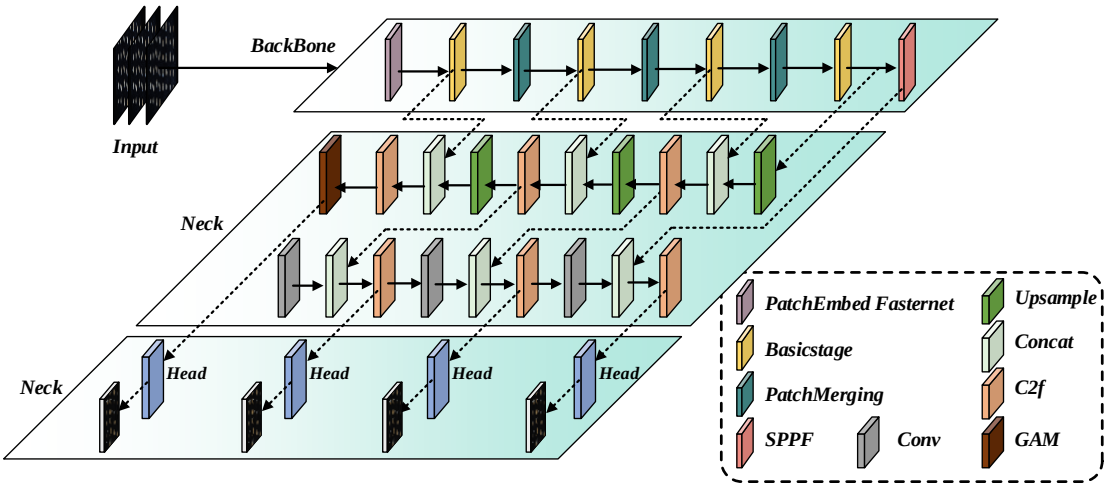

Fig. 1. Transformer model based on multimodal.

As shown in Fig. 1, the model takes text mode as the main component, speech mode, and image mode as the secondary components, and the characteristics of the three modes are respectively input into the multi-head attention module. Through this setting, the text data containing rich information can guide the voice and image data, which can be used to mine multi-modal association information. The process of calculating the text feature attention score with this module is as follows:

Firstly, the text features are divided into three vectors: query vector Ql, keyword vector Kl and true value vector Vl, and all the vectors are linearly transformed; Then, Ql and Kl are sent to calculate the attention score, and the dimension Kl is used to limit the calculation result to ensure that the inner product is not too large; Finally, the final calculation result is obtained by weighted summation of attention score and Vl. Specifically, as shown in Formula (1).

$$Attention(Q_l, K_l, V_l) = \text{softmax}((Q_l K_l^T)/\sqrt{d_k})V_l \quad (1)$$

The above calculation process is performed multiple times, and each calculation is regarded as a head [11, 12]. By splicing the results of multiple heads, the final multi-head attention calculation result can be obtained, as shown in Formulas (2) and (3).

$$head_i = Attention(Q_l W^Q, K_l W^K, V_l W^V) \quad (2)$$

$$F_{(l)} = MHA((Q_l, K_l, V_l)) = Concat(head_1, ..., head_h)W^O \quad (3)$$

After getting the calculation result of attention, it is passed into FFN to mine the nonlinear relationship of features, so as to enhance the performance ability of features, as shown in Formula (4).

$$FFN = Relu(H'W^1 + b^1)W^2 + b^2 \quad (4)$$

Each layer in the LGT needs to be processed, as shown in Formula (5).

$$F(x) = LayerNorm(x + Sublayer(x)) \quad (5)$$

For minor components such as speech features and image features, the query vector comes from the text mode, and the keyword vector and the true value vector come from the speech and image modes when calculating the multi-head attention [13, 14]. When processing speech and image features, text features are used to introduce information from different representation spaces, as shown in Formulas (6) and (7).

$$F_{(a)} = MHA(Q_l, K_a, V_a) = Concat(head_1, ..., head_h)W^O \quad (6)$$

$$F_{(v)} = MHA(Q_l, K_v, V_v) = Concat(head_1, ..., head_h)W^O \quad (7)$$

B. Soft Mapping Module

The model has learned the interaction information between the modes and needs to project the learned results of each mode into a new performance space in the soft mapping module for fusion before classification [15]. Precisely, the results output by the forward propagation network are first mapped to a higher-dimensional space, as shown in Formula (8).

$$NewMatrix = W_m M \quad (8)$$

Then the soft attention is calculated for each matrix in the high-dimensional space, and then the weighted sum of the results is integrated into the vector to obtain the calculation result of soft attention [16, 17]. This calculation process is shown in Formulas (9) and (10).

$$p_i = softmax((v_i^p)^T (NewMatrix)) \quad (9)$$

$$SoftAttention_i(M) = m_i = \sum_{j=0}^{N} (p_{ij} M_j) \quad (10)$$

Finally, after stacking these results, you can get the results of Soft Mapping, as shown in Formula (11).

$$s = Stacking(\sum_{j=0}^{N}(m_j)) \quad (11)$$

Note that a residual calculation and Layer Normalization are performed at the end of this process to ensure that the next round of input includes the results of the previous round, as shown in Formula (12).

$$M = LayerNorm(M + s) \quad (12)$$

The result s obtained above is the result after processing the respective output matrix M of each mode, and the vectors obtained by each mode are summed in order of elements, and the summed results are classified and predicted according to Formula (13).

$$y \sim p = W_p(LayerNorm(s_l + s_a + s_v)) \quad (13)$$

C. Construction of Recommendation Model Fusing Interaction of Bert and High-order Dominant Features

1) The overall structure of model fusing Bert interaction with high-order dominant features: In this paper, according to the actual situation of video recommendation, the Bert model is integrated into the X Deep FM framework. This model can extract text feature vectors with deeper semantics through the Bert model and obtain their text feature vectors by extracting text information such as video titles and tags Because categorical discrete features such as user ID, video ID, and related attributes are difficult to directly use as inputs to deep learning models [18, 19]. Therefore, Label Encoder is used to convert into categorical codes, discontinuous values or texts are converted into categorical codes, and then Embedding is used to convert them into low-dimensional, dense feature vectors, which are input into the model. Finally, the input feature vector and the user's preference degree value are used to iteratively update the training model to improve accuracy and reliability [20]. The network structure of its overall model is shown in Fig. 2.

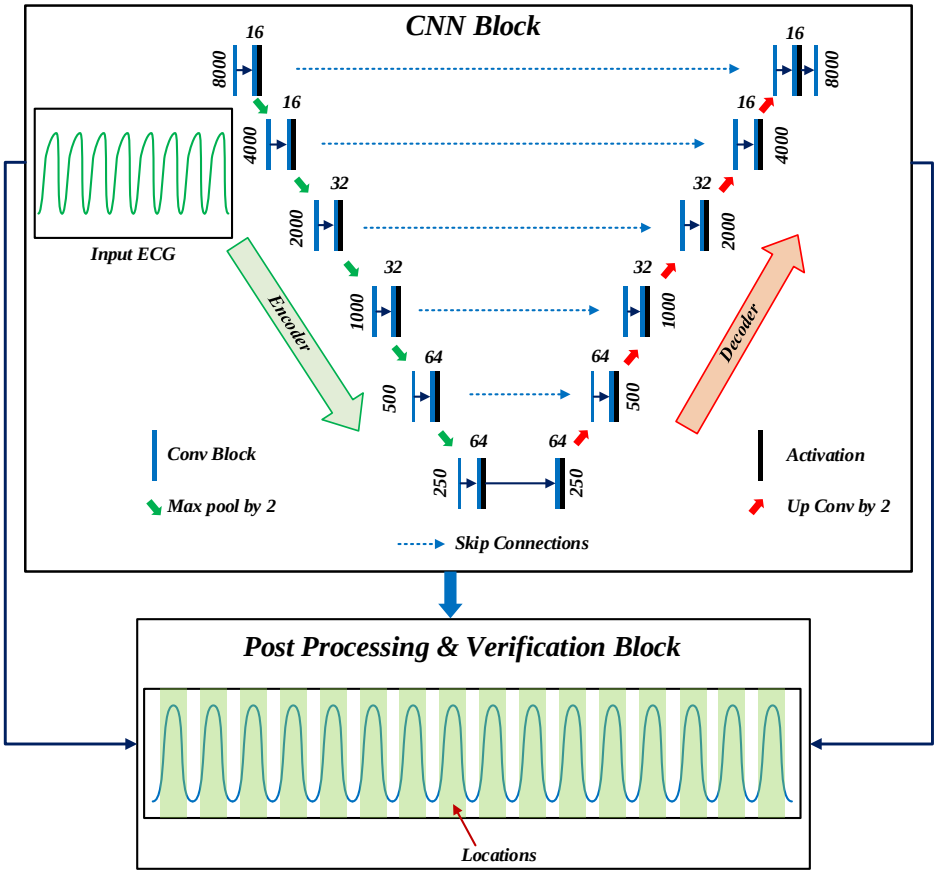

Fig. 2. Fusion model of Bert interaction with high-order explicit features.

The Fig. 2 shows a model that combines Bert and high-order explicit features. It is an end-to-end recommendation model, which has the characteristics of both low-order and high-order feature interactions and implicit and explicit features. The overall structure of the network is composed of four parts: input and text feature extraction, compressed interactive network part, multi-layer neural network and score prediction (that is, output layer), in which the compressed interactive network can extract high-order explicit cross features of the input layer, and the multi-layer neural network can extract high-order invisible cross features [21, 22].

2) Extraction of text information features by Bert: Bert is a bidirectional Transformer-based encoder, which is a bidirectional model obtained by unsupervised training on large-scale corpus. Compared with the GPT model, the Bert model

uses the Encoder structure in Transformer as the main component of the model. The Bert model reads the text, constructs special characters in the text, completes the training through the multi-layer Transformer Encoder [23, 24] structure, and finally uses the vector corresponding to the special characters as the output of the Bert model. In the training task of the language model, the special characters are constructed in the form of filling. By randomly setting [MASK] on the input, let the model predict the words at this position to complete the training of the model. In actual use, it splices special characters [CLS] on the text, so that the output of special characters is matched with the target task to achieve. At present, the Bert model has achieved remarkable results in extractive tasks (SQuAD), sequence labeling tasks (named entity recognition), and classification tasks (SWAG) and other tasks.

3) Input layer: The input layer is responsible for transforming users and characteristics into the form required by the model so that input information can be better understood and processed. In the process of processing features, when using categorical discrete features, first use Label Encoder to encode, the value is between 0 and n-1, so that this feature can be recognized by the model. For numerical continuous features (such as playback volume, user level, etc.), they can be directly entered into the model as input features for calculation [25]. For text-like features, we use the previous Bert model to extract sentence vectors. Since the dimension of each sentence vector is 768 dimensions, all features are spliced together to form the final input vector.

The combination of the sub-types and continuous numerical types processed by the above three methods will cause problems such as dimension explosion and excessive resource occupation, and it is not very good for neural networks to deal with this input. In this paper, the Embedding layer is used to deal with subtyping and continuous numeric types, so as to solve the problems of dimension explosion and excessive resource occupation [26, 27]. By this method, the original sparse matrix is transformed into a dense continuous vector with suitable length, so that the neural network can better handle this input. Although the initial feature length of the sample data may vary, Embedding can still effectively improve this situation, thereby increasing the accuracy and reliability of the model. After the feature embedding layer processing, its length will remain unchanged and will not be affected by the outside world. After this process, follow-up deep learning operations are carried out. The network structure essentially forms a weight matrix. According to a certain mapping relationship, the weight information of the original matrix is transformed into a new dimension matrix through matrix multiplication calculation. According to the reverse mapping relationship, the matrix is multiplied, and the original matrix will be restored matrix. The application of Embedding layer can effectively reduce the sparsity of data, and can change the original isolated vectors into closely related vectors, which can greatly enhance the scalability of the algorithm.

4) DNN layer: To deeply explore the feature interaction relationships implicit in the information, deep neural networks (DNNs) are used for learning. DNN is developed from the multi-layer perceptron (MLP) technology, which has deeper network layers and more types of activation functions. It can connect multiple hidden layers of nonlinear structures, fitting complex function curves and mining deeper interaction features through large-scale training data [28]. The deep neural network performs excellently, can deal with complex problems, and has remarkable effects. Its powerful ability is mainly due to the large number of neural network layers; that is to say, the more network layers, the more complex and in-depth the neural network, and the more learning. Accurate. The basic structure of the neural network comprises three parts: the input layer, the hidden layer, and the output layer. The connection mode between layers is a complete connection, and there is at least one hidden layer. The more hidden layers there are, the higher the expressive ability of the model. The relationship between layers of the deep neural network is nonlinear, and the task of the lower network layer is to extract low-order edge features with relatively simple relationships from the original input data. Each neuron in the bottom network layer acquires some low-order information. More advanced local features can be obtained by combining the underlying information on the middle-hidden layer. The top layer fuses local features into higher-level features. However, it is impossible to theoretically understand the crossover characteristics of each DNN neural network layer and the characteristics each neuron represents. After the training, which feature interactions are more effective in the entire neural network cannot be explained, so these unexplained high-order feature interactions are considered implicit feature interactions. However, the experimental results confirm that DNN can unearth unintelligible but effective high-order feature interactions, which are called implicit feature interactions, and the scattered results are shown in Fig. 3.

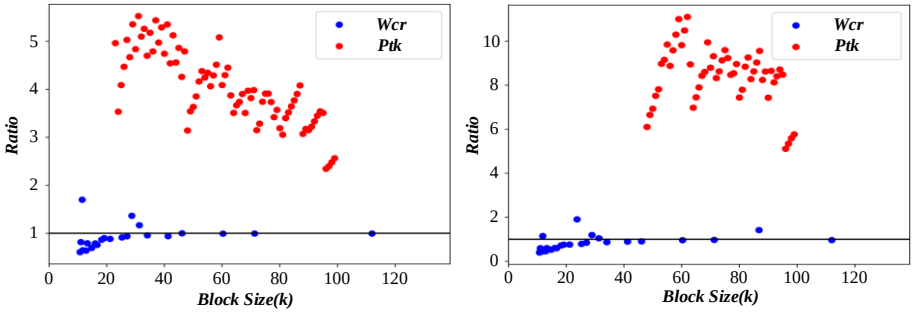

Fig. 3. Interactive dispersion results of implicit features.

III. Design of Cross-Border E-Commerce Short Video Recommendation System based on Multi-task Learning

A. Multi-level Association Mining Framework

In exploring low-level correlations, this study adopts a model-agnostic approach to fuse features, aiming to uncover the intricate feature associations spanning across multi-modal representations [29]. Drawing inspiration from tensor fusion networks, we introduce the use of Unfolded Fusion Functions (UFF) to address the limitations of traditional fusion techniques, such as multimodal splicing. By leveraging UFF, we elevate single-modal features into higher-dimensional spaces, facilitating their fusion. This approach employs a 3-fold Cartesian product to seamlessly integrate multiple single-modal representations, capturing both bimodal and trimodal interactions through a multi-level fusion process.

$$\{(T^l, T^a, T^v)/T^l \in [{}_1^{T^l}], T^a \in [{}_1^{T^a}], T^v \in [{}_1^{T^v}]\} \tag{14}$$

$$F_{(m)} = [{}_1^{T'}] \otimes [{}_1^{T^a}] \otimes [{}_1^{T^v}] \tag{15}$$

The precise computational methodology is outlined in Formulas (14) and (15), offering a nuanced and robust framework for analyzing and utilizing multi-modal data.

B. Multi-task Learning Framework

In this section, we introduce a multi-task learning framework that is designed to tackle diverse analysis tasks through the employment of a rigorous hard parameter sharing mechanism. This mechanism enables all tasks to synergistically harness neurons and weights in the foundational low-level network, while reserving task-specific neurons and weights for each individual task in the higher-level network. The framework adopts a two-tiered architecture, with the underlying representation learning network serving as a common layer and the prediction network tailored to meet the distinct demands of each task.

$$F_s^* = ReLU(F_s W_s^{1T} + b_s^1) \tag{16}$$

$$y_s = F_s^* W_s^{2T} + b_s^2 \tag{17}$$

Within this multi-task learning paradigm, four distinct tasks are formulated, and the specific configuration of the task-oriented layers is outlined in Formulas (16) and (17). Notably, the single-modal task is trained utilizing labels generated by the SLGM methodology, limiting its existence to the training phase. Ultimately, the model relies on the predicted outcomes of the multi-modal task as the definitive output, reflecting its emphasis on the integration of multimodal information. This approach offers a comprehensive and efficient solution for multi-task learning, promoting knowledge sharing and task specialization within a unified framework.

C. Self-Supervised Label Generator

Most multimodal analysis datasets need more independent single-modal labels, posing a challenge for multi-task learning frameworks. To address this limitation, Fig. 4 illustrates the outcomes of a self-supervised label generation module tailored to diverse modalities [30]. Consequently, this section introduces the Self-Supervised Label Generation Module (SLGM), whose primary objective is to derive single-modal annotations from multimodal annotations. The conceptual foundation of SLGM is grounded in two potential mapping relationships: (1) a direct correlation between modal representations and their corresponding modal supervision values and (2) a proportionality in the mapping relationships among different modalities. By harnessing these insights, SLGM aims to bridge the gap between multimodal annotations and the desired single-modal labels, enabling a more comprehensive and practical multi-task learning framework.

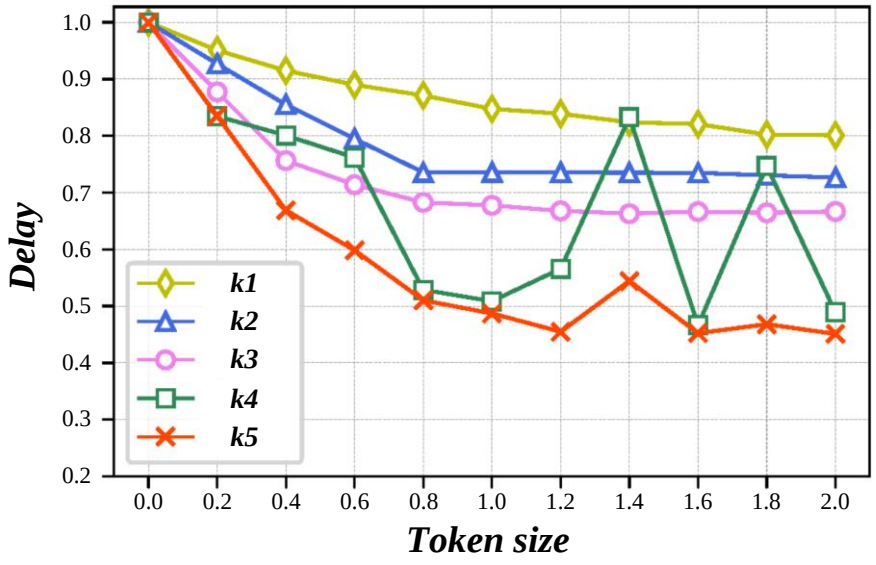

Fig. 4. Results of self-supervised label generation module in different modes.

The details is shown in Formula (18). The (SLGM delineates modal representations into two distinct categories, dictated by their polarity. Subsequently, it identifies the central tendencies of these two categories, yielding a modal representation positive center and a modal representation negative center for each modality. The precise computational methodology for this categorization and centering is outlined in Formulas (19) and (20), ensuring a rigorous and systematic approach to generating single-modal annotations from multimodal data.

$$C = (F_m \# L_m) \propto (F_u \# L_u) \tag{18}$$

$$C_p = \frac{\sum_{i=1}^{N} I(y(i) > 0) \cdot F_i}{\sum_{i=1}^{N} I(y(i) > 0)} \tag{19}$$

$$C_n = \frac{\sum_{i=1}^{N} I(y(i) < 0) \cdot F_i}{\sum_{i=1}^{N} I(y(i) < 0)} \tag{20}$$

Next, the SLG) employs the coefficient as a metric to quantify the degree of deviation between each sample and its corresponding class center. This calculation is precisely defined in Formulas (21) and (22), providing a rigorous mathematical framework for assessing the proximity of samples to their respective modal representation centers.

$$S_p = \sum_{j=1}^{K} \sqrt{F(j) C_p(j)} \tag{21}$$

$$S_n = \sum_{j=1}^{K} \sqrt{F(j) C_n(j)} \tag{22}$$

IV. Experimental Results and Analysis

A. Quantitative Analysis

The model was quantitatively analyzed using CMU-MOSI and CMU-MOSEI datasets. During the test, the multimodal analysis task was regarded as a regression and classification task, respectively. The regression task used mean absolute error (MAE) and F1 score as evaluation indicators. Among them, the smaller the value of MAE, the better the model performance, and the larger the value of other indicators, the better the model performance. As shown in Table I and Table II, the five average experimental results on the two data sets show that the proposed model can achieve satisfactory results and its performance can reach the average level. This confirms that the model can effectively mine multimodal associations to improve prediction effects.

TABLE I. Experimental Results of the Model On CMU-MOSI Dataset

Model	MAE	F1-Score
MFN	0.95	78.1
RAVEN	0.92	76.6
MCTN	0.91	79.1
MulT	0.87	82.8
MISA	0.78	83.6
Self _ MM	0.71	86.0
Ours	0.81	82.9

TABLE II. Experimental Results of the Model on CMU-MOSEI Dataset

Model	MAE	F1-Score
MFN	0.71	77.0
RAVEN	0.61	79.5
MCTN	0.61	80.6
MulT	0.58	82.3
MISA	0.55	85.3
Self _ MM	0.53	85.3
Ours	0.59	82.2

Compared with other models, there are still some small gaps in some indicators. As can be seen from the analysis of the reasons in Fig. 5, the MISA model and the Self _ MM model have already processed the data in the representation learning stage, and improved the quality of modal representation by learning the common and individual information of different modal data, which means that such models More reliable data can be obtained at the beginning to improve the subsequent prediction effect. During the experiment, more than 4,000 samples were randomly selected from the test set to test the 2 classification results. The plotted curves are shown in Fig. 6, which can reflect the excellent performance of the model.

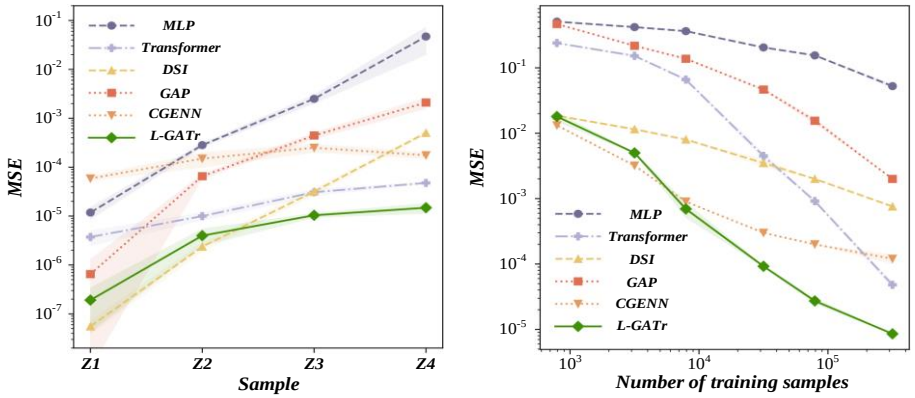

Fig. 5. Results of the MISA model and the Self-MM model in the presentation learning stage.

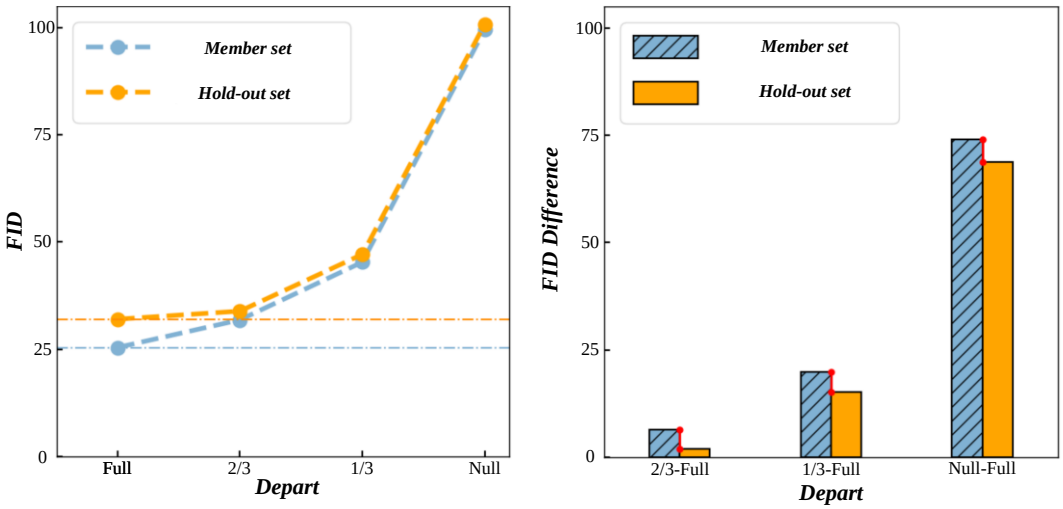

Fig. 6. Performance differences between models.

Table III shows the results from the CMU-MOSI dataset. In order to compare the performance differences between models, this chapter chooses the evaluation indicators of the classification task and the regression task for comparison. For classification tasks, the text-proposed model has obvious advantages in the evaluation of 2 classification accuracy (Acc-2), 7 classification accuracy (Acc-7) and F1 score. For the regression task, the model has achieved significant improvements in the evaluation of both the mean absolute error MAE and the Pearson correlation coefficient Corr, and the results are shown in Fig. 6. In addition to the MAE indicator, the larger the evaluation value shown in the table, the better the performance of the model on this indicator. Fig. 7 is the index result graph of the linear level. The experimental results show that the multi-modal analysis using the multi-task learning framework provides a new idea to solve the problems in this field. The model performance with the help of multi-task joint training is better than that with a single task. Task-trained model performance. In addition, the multi-level association mining framework also proves its effectiveness, it can obtain more useful information than single-angle mining.

TABLE III. EXPERIMENTAL RESULTS OF THE MODEL ON THE CMU-MOSI DATASET

Model	MAE	Corr	Acc-7	Acc-2	F1-Score
MFN	0.95	0.66	36.2	78.1	78.1
RAVEN	0.92	0.69	33.2	78.0	76.6
MCTN	0.91	0.68	35.6	79.3	79.1
MulT	0.87	0.70	40.0	83.0	82.8
MISA	0.78	0.76	42.3	83.4	83.6
Self _ MM	0.71	0.80	46.7	86.0	86.0
Ours	0.69	0.81	47.1	88.4	88.4

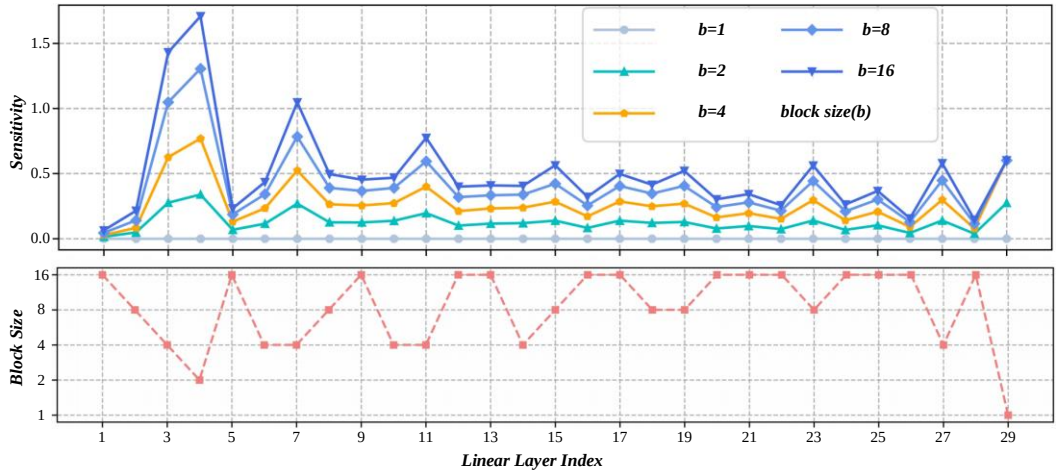

Fig. 7. Linear layer index result graph.

In this paper, a Transformer-based multi-analysis method is proposed. The model can fully consider the relationship between multiple information, use the linguistic-guided Transformer to mine the association between multiple data, and use the soft mapping module to achieve tight coupling of multiple data, thereby improving the analysis effect of the model. The experiments of the model on two data sets have achieved satisfactory results, which further demonstrates the feasibility

and effectiveness of the theory of improving the prediction effect by mining multiple associations and providing a specific solution for the problems in multiple analysis fields. Table IV shows results from the dataset, which shows that the model can also achieve good results.

TABLE IV. EXPERIMENTAL RESULTS OF THE MODEL ON THE CMU-MOSEI DATASET

Model	MAE	Corr	Acc-7	Acc-2	F1-Score
MFN	0.71	0.54	45.0	76.9	77.0
RAVEN	0.61	0.66	50.0	79.1	79.5
MCTN	0.61	0.67	49.6	79.8	80.6
MulT	0.58	0.70	51.8	82.5	82.3
MISA	0.56	0.76	52.2	85.5	85.3
Self _ MM	0.53	0.77	52.4	85.2	85.3
Ours	0.51	0.74	53.9	86.2	85.9

B. Ablation Experiment

The proposed model includes two structures: LGT and SM. The former interacts between modes to improve the learning effect when learning a certain mode, and the latter maps the learning results of each mode to a high-dimensional space for better classification. In order to verify effectiveness of two structures, this section conducts ablation experiments on the CMU-MOSI dataset, which are specifically divided into four situations: LGT and SM are not used at all; only remove LGT; Only SM is removed; LGT and SM were used simultaneously. Choosing to use the ordinary multi-head attention mechanism instead of LGT when it is not used means that the interaction ability between modes is lost. When SM is not used, the results of independent learning of each modal are directly weighted and averaged, and then classified. The ablation experimental results are shown in Fig. 8 and Fig. 9, from which it can be seen that the two main structures of the model can play a positive role in the final prediction.

C. Validation Experiment of Self-Supervised Label Generator

Combining the idea of multi-task learning with the proposed model, a multimodal analysis model based on a multi-level association mining framework and a self-supervised label generator is proposed in this chapter to solve the multimodal analysis problems faced in multimodal analysis simultaneously. Modal association mining and multimodal collaborative learning problems. The multi-level association mining framework further deepens the research content, which can simultaneously mine association information from two angles. The self-supervised tag generator can automatically train the single-modal tag-assisted multi-task learning framework, thus realizing multimodal collaborative learning. The verification experiment of the self-supervised label generator is shown in Fig. 10. A large number of experiments have been carried out on classical data sets in the field of multimodal analysis, all of which prove that the proposed model has excellent analytical performance and can provide a feasible idea for solving the problems existing in this field.

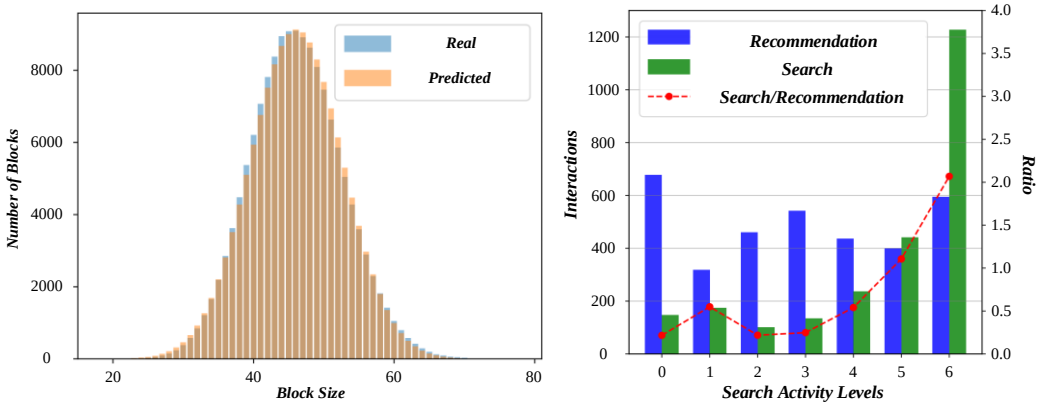

Fig. 8. Ablation experiment of module size and search level.

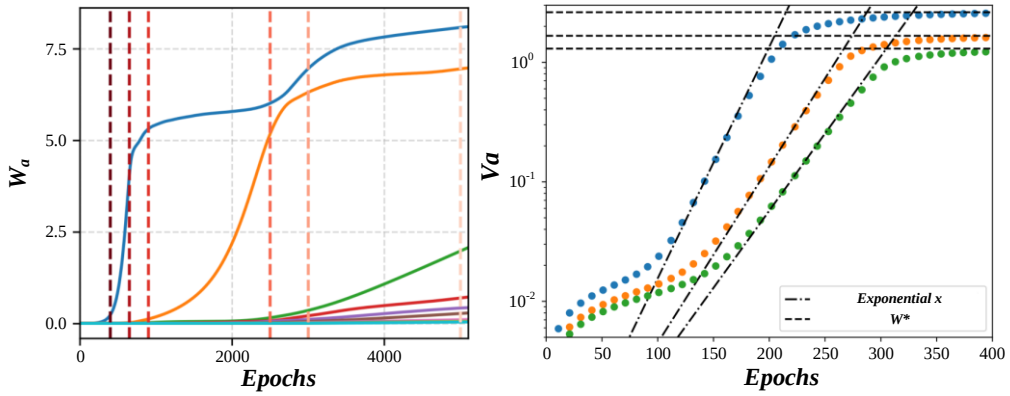

Fig. 9. Result diagram under different epoch.

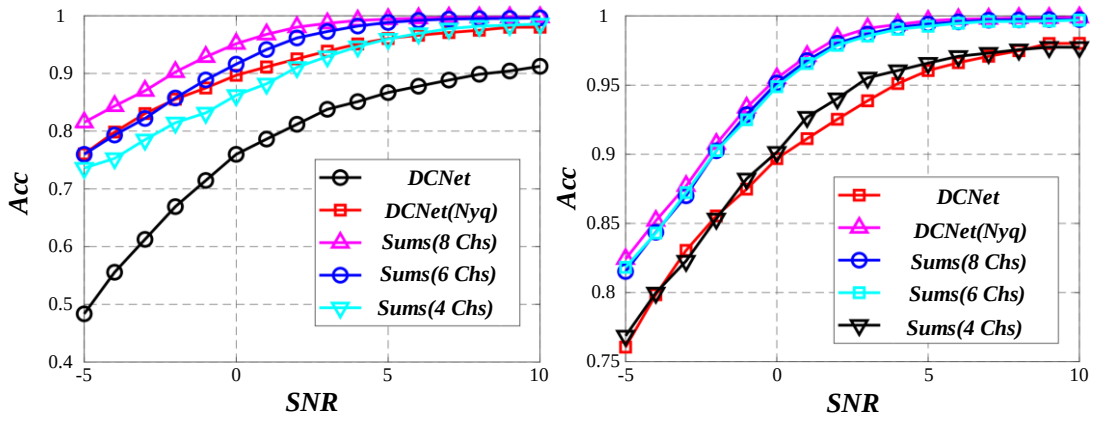

Fig. 10. Verification experiment of self-supervised label generator.

In order to examine the rationality and robustness of SLGM, this section extracts several labels generated by SLGM during training. As shown in Table V, the polarity of the single-modal labels of samples 1 to 3 is consistent with that of the manually labeled multi-modal labels, which shows that the single-modal labels generated by SLGM are of some value. Compared with the manual labeling of multi-modal labels, the single-modal labels of samples 4 to 5 achieve a negative shift, and this negative shift is reasonable.

TABLE V. SAMPLE SLGM GENERATION LABEL SAMPLE

MOSI-M	SLGM-L	SLGM-A	SLGM-V
2.2	1.1925	0.3874	0.8557
2.4	1.7874	0.0639	1.1984
-1.8	-1.5281	-0.9052	-1.3683
-0.2	-0.0327	0.0874	0.0001
0.6	0.8281	0.0052	-0.0856

V. SUMMARY

This research has conducted an in-depth exploration of the design of a cross-border e-commerce short video recommendation system based on the multi-modal fusion Transformer model. An efficient and accurate recommendation system has been successfully built through the comprehensive use of advanced technologies such as Transformer and multi-tasking learning. Many short videos and user behavior data of cross-border e-commerce platforms were collected and analyzed during the research process, supporting system design and optimization. First of all, in the aspect of multi-modal association mining, the cross-modal multi-head attention model is used to conduct in-depth analysis of multi-modal data such as video and text, and it is found that there is rich association information between different modes. A multi-modal analysis model is designed based on the multi-task learning framework in multi-modal collaborative learning. The model can learn more comprehensive and accurate multi-modal data representation through modal information sharing during training. In addition, the self-supervised labeling method proposed in this paper effectively solves the problem of missing labels, making the model perform well under limited labels. At the sorting level, we fuse the Bert and high-order explicit feature cross models to extract deeper text features and capture the deep-seated interaction relationship between users and short videos. Through the analysis and verification of large-scale data, this feature fusion method can significantly improve the prediction accuracy of the recommendation system and provide users with more accurate recommendation services. Finally, a complete cross-border e-commerce short video recommendation system is designed and implemented. The system integrates multiple modules such as data acquisition, preprocessing, recall, and sorting and realizes the stable operation of the system and user-friendly interaction through front-end and back-end design. In practical application, the system has effectively improved user satisfaction and recommendation efficiency, with an accuracy rate and F1 score of 98.76% and 97.89%, respectively, bringing significant value to cross-border e-commerce platforms.

REFERENCES

[1] Gandhudi, M., Alphonse, P. J. A., Velayudham, V., Nagineni, L., & Gangadharan, G. R. (2024). Explainable causal variational autoencoders based equivariant graph neural networks for analyzing the consumer purchase behavior in E-commerce. Engineering Applications of Artificial Intelligence, 136, 108988.

[2] 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.

[3] Zeng, F. (2023). Multimodal music emotion recognition method based on multi data fusion. International Journal of Arts and Technology, 14(4), 271-282.

[4] Dutta, M., & Ganguly, A. (2024). Incremental-based YoloV3 model with Hyper-parameter Optimization for Product Image Classification in E-commerce Sector. Applied Soft Computing, 112029.

[5] Gwak, M., Cha, J., Yoon, H., Kang, D., & An, D. (2024). Lightweight Transformer Model for Mobile Application Classification. Sensors, 24(2).

[6] Gai, T., Wu, J., Liang, C., Cao, M., & Zhang, Z. (2024). A quality function deployment model by social network and group decision making: Application to product design of e-commerce platforms. Engineering Applications of Artificial Intelligence, 133, 108509.

[7] Zhuang, S. (2024). E-commerce consumer privacy protection and immersive business experience simulation based on intrusion detection algorithms. Entertainment Computing, 100747.

[8] Zhao, H. (2021). A Cross-Border E-Commerce Approach Based on Blockchain Technology. Mobile Information Systems, 2021.

[9] Zhang, C., Zheng, H., & Wang, Q. (2022). Driving Factors and Moderating Effects Behind Citizen Engagement with Mobile Short-Form Videos. Ieee Access, 10, 40999-41009.

[10] Chang, C., Zhou, J., Weng, Y., Zeng, X., Wu, Z., Wang, C.-D., & Tang, Y. (2023). KGTN: Knowledge Graph Transformer Network for explainable multi-category item recommendation. Knowledge-Based Systems, 278.

[11] Gu, P., Hu, H., & Xu, G. (2024). Modeling multi-behavior sequence via HyperGRU contrastive network for micro-video recommendation. Knowledge-Based Systems, 295, 111841.

[12] Zhu, H., Wei, H., & Wei, J. (2023). Understanding users' information dissemination behaviors on Douyin, a short video mobile application in China. Multimedia Tools and Applications.

[13] Li, C., Cao, Y., Zhu, Y., Cheng, D., Li, C., & Morimoto, Y. (2024). Ripple Knowledge Graph Convolutional Networks for Recommendation Systems. Machine Intelligence Research, 21(3), 481-494.

[14] Li, P., Li, T., Wang, X., Zhang, S., Jiang, Y., & Tang, Y. (2022). Scholar Recommendation Based on High-Order Propagation of Knowledge Graphs. International Journal on Semantic Web and Information Systems, 18(1).

[15] Jing, H. (2022). Application of Improved K-Means Algorithm in Collaborative Recommendation System. Journal of Applied Mathematics, 2022.

[16] Shen, X. (2023). E-commerce User Recommendation Algorithm Based on Social Relationship Characteristics and Improved K-Means Algorithm. International Journal of Computational Intelligence Systems, 16(1).

[17] Du, H., Tang, Y., & Cheng, Z. (2023). An efficient joint framework for interacting knowledge graph and item recommendation. Knowledge and Information Systems, 65(4), 1685-1712.

[18] Zhang, L., Zhang, W., McNeil, M. J., Chengwang, N., Matteson, D. S., & Bogdanov, P. (2021). AURORA: A Unified fRamework fOR Anomaly detection on multivariate time series. Data Mining and Knowledge Discovery, 35(5), 1882-1905.

[19] Matrouk, K. M., Nalavade, J. E., Alhasen, S., Chavan, M., & Verma, N. (2023). MapReduce Framework Based Sequential Association Rule Mining with Deep Learning Enabled Classification in Retail Scenario. Cybernetics and Systems.

[20] Chen, Z., & Ge, Z. (2022). Knowledge Automation Through Graph Mining, Convolution, and Explanation Framework: A Soft Sensor Practice. Ieee Transactions on Industrial Informatics, 18(9), 6068-6078.

[21] Huu-Thiet, N., Li, S., & Cheah, C. C. (2022). A Layer-Wise Theoretical Framework for Deep Learning of Convolutional Neural Networks. Ieee Access, 10, 14270-14287.

[22] Eun, Y. J., Chae, S., & Kyungmin, B. (2022). Layered Abstraction Technique for Effective Formal Verification of Deep Neural Networks. Journal of KIISE, 49(11), 958-971.

[23] Wu, W., Wang, W., Jia, X., & Feng, X. (2024). Transformer Autoencoder for K-means Efficient clustering. Engineering Applications of Artificial Intelligence, 133.

[24] Lo, P.-C., & Lim, E.-P. (2023). A transformer framework for generating context-aware knowledge graph paths. Applied Intelligence, 53(20), 23740-23767.

[25] Feng, Y., Zhai, M., & Du, Y. (2024). The effects of mini-detail short videos on consumer purchase intention on Taobao: A TAM2-based approach. Entertainment Computing, 100745.

[26] Wang, C., & Xiao, Z. (2022). A Deep Learning Approach for Credit Scoring Using Feature Embedded Transformer. Applied Sciences-Basel, 12(21).

[27] Fan, J., Huang, L., Gong, C., You, Y., Gan, M., & Wang, Z. (2024). KMT-PLL: K-Means Cross-Attention Transformer for Partial Label Learning. Ieee Transactions on Neural Networks and Learning Systems.

[28] Anitha, J., & Kalaiarasu, M. (2022). A new hybrid deep learning-based phishing detection system using MCS-DNN classifier. Neural Computing & Applications, 34(8), 5867-5882.

[29] Xu, R., Li, J., Li, G., Pan, P., Zhou, Q., & Wang, C. (2022). SDNN: Symmetric deep neural networks with lateral connections for recommender systems. Information Sciences, 595, 217-230.

[30] Nam, W., & Jang, B. (2024). A survey on multimodal bidirectional machine learning translation of image and natural language processing. Expert Systems with Applications, 235.