Transformer Model Optimization Method for Multi-Modal Data Fusion

Shanshan Yang*, Jie peng
College of Information Engineering, Jiaozuo University, Jiaozuo 454000, China

***Abstract*—This study proposes an optimized Transformer model for multimodal data fusion tasks, designed to address the challenges of data fusion from different modes such as text, image, and audio. By improving data preprocessing methods, optimizing model architecture and fusion strategies, the study significantly improves the performance of the model in multimodal tasks. The experimental results show that the optimized model is superior to the benchmark model and other comparison models in key indicators such as accuracy, recall, F1 score and AUC value, and shows stronger performance and higher stability. In particular, the research solves the problems of data heterogeneity and computing resource consumption by introducing a weighted fusion strategy, multi-head self-attention mechanism and lightweight design. At the same time, the processing of missing modal data is optimized to enhance the robustness of the model. Despite the remarkable results, there are still challenges such as data heterogeneity, computational efficiency, and missing modal data. Future research can further optimize modal alignment methods and data preprocessing techniques to improve the performance of the model in practical applications. This research provides a new idea and direction for the application and development of multimodal data fusion technology.**

***Keywords*—Transformer model; multimodal data fusion; model optimization; attention mechanism; adaptive fusion**

I. INTRODUCTION

Multimodal data fusion is a research hotspot in the field of artificial intelligence, especially in natural language processing, computer vision, speech recognition and other application fields. The traditional single-modal learning method can only deal with a certain type of data, while the multi-modal data fusion method can improve the generalization ability and prediction accuracy of the model by integrating data from different sources. With the development of deep learning technology, Transformer model has become a core method in multi-modal data fusion due to its excellent performance in sequence modeling tasks. Transformer architecture can effectively handle large scale data and capture long-term dependencies, providing better performance when multiple modal data is merged. The complex multi-modal data Transformer model still has some limitations, such as large consumption of computing resources and low training efficiency. Therefore, how to optimize the Transformer model to make it efficient for multi-modal data fusion is the challenge of current research. In view of these problems, this study will explore effective methods to optimize Transformer model to improve its application effect in multi-modal data fusion.

*Corresponding Author.

The Transformer model in the field of multimodal data fusion is widely used because of its excellent performance in processing complex data sequences. Wu et al. proposed that cross-modal learning can be used to optimize the local feature representation of vision-thermal infrared character recognition, thereby enhancing Transformer's identification capability in cross-modal tasks [1]. Xue et al. discussed the importance of multi-modal data consistency in fake news detection, indicating that the integration of multiple data sources is conducive to improving the accuracy of fake news identification [2]. Jing et al. realizes multi-modal fake news detection through a progressive fusion network, demonstrating Transformer's advantages in information fusion [3]. De Melo is a key reference for the application of the Transformer model in the field of multimodal data fusion, and provides the research direction of Transformer structure optimization. This literature discusses the challenges of the Transformer in processing multimodal data, especially how to improve the fusion effect through the self-attention mechanism [4]. Ghorbanali et al. used weighted convolutional neural networks for sentiment analysis and integrated transfer learning strategies to improve the accuracy of multimodal sentiment analysis [5]. Golovanevsky et al. proposed a deep learning method based on attention mechanism for the diagnosis of Alzheimer's disease, and achieved good results in multi-modal medical data fusion [6]. Chango et al. studied the method of using multi-modal data to predict student performance in the field of education, and proposed an optimization method based on attribute selection and multi-modal data integration, which can effectively improve the prediction accuracy [7]. Bao et al. design a reward-based crowdfunding success prediction framework based on multi-modal data, and optimize the performance of Transformer model in data fusion with a theory-driven approach [8]. Lin and Hu proposed the MissModal method to enhance the robustness of missing modes in multi-modal sentiment analysis, thereby improving the performance of Transformer model in the context of missing data [9]. Ciroku et al. studied ontology-based multi-modal data integration methods for automating sensing and inference tasks, and emphasized the importance of multi-modal data in semantic integration [10].

Although existing research has made progress in multimodal data fusion and Transformer optimization, there are still several key problems that require further investigation. These will be discussed in detail in the next section.

Several solutions proposed by researchers have been discussed, such as weighted fusion strategies, hierarchical attention mechanisms, and adaptive alignment methods. These

studies demonstrate the potential of Transformer-based architectures for addressing data heterogeneity and improving feature integration. However, existing solutions still face challenges, including high computational resource requirements and limited robustness when handling missing modality data. Compared to these approaches, this research focuses on optimizing Transformer architecture through lightweight design and dynamic fusion strategies, aiming to enhance both performance and efficiency. This positioning highlights the novelty of the proposed method and its contribution to advancing multimodal data fusion techniques.

These studies show that while multimodal data fusion has made positive progress in various applications, Transformer model optimization is still the key to improving its performance. Various studies have proposed different optimization methods, including model structure adjustment, data fusion strategy and robustness improvement of missing data, which provide valuable references for subsequent research and application.

Although the Transformer model has made significant progress in handling single-modal data, it still faces many challenges in multi-modal data fusion tasks. Firstly, the heterogeneity of multi-modal data brings difficulties to data preprocessing and feature extraction. Data of different modes have different scales, noise levels and redundancy, which puts forward higher requirements for model training. Secondly, the existing Transformer model has a huge computing overhead in the face of large-scale multi-modal data, resulting in long training time and high resource consumption. In addition, the existing models have poor robustness when dealing with missing mode data, and lack effective strategies to deal with the missing problem between different modes. Finally, how to improve the interpretation and scalability of the model while ensuring its performance is also an important difficulty in the current research.

This study aims to propose an optimized Transformer model for multi-modal data fusion tasks. The main goal of this research is to improve the Transformer's performance in processing multi-modal data by improving data preprocessing methods, model structure and fusion strategies. Specifically, the research will focus on solving the problems of computational efficiency, robustness and generalization ability of models in multimodal data fusion. By exploring effective data fusion strategies, this study will also explore how to deal with missing modal data to enhance the adaptability and reliability of the model in practical applications. In addition, this research also plans to optimize the existing Transformer architecture to improve its performance in multimodal tasks and promote the application range and accuracy of multimodal data fusion technology.

This study will adopt a combination of theoretical analysis and experimental verification. Firstly, a systematic literature review and theoretical analysis of the existing multi-modal data fusion methods of Transformer model will be conducted to clarify the advantages and disadvantages of current technologies. Then, on this basis, an improved Transformer model is designed to optimize data pre-processing, feature extraction and fusion strategies. Specifically, the research will enhance the correlation between different modal data by introducing a multi-level feature selection mechanism, thereby improving the fusion efficiency of the model. At the same time, simulation data and actual data will be used for model training and evaluation, and multiple indicators such as accuracy rate and recall rate will be used to comprehensively evaluate the performance of the model. To solve the problem of missing modal data, a self-supervised learning mechanism will be introduced to improve the robustness and stability of the model.

This research has important theoretical value and practical significance in the field of multimodal data fusion. From a theoretical perspective, the proposed optimization method will promote the application of Transformer model in multi-modal data fusion, especially in the aspects of data heterogeneity, missing modes and computational efficiency. This can not only enrich the existing multimodal learning theory, but also provide new ideas and methods for the subsequent research. From the perspective of practical application, the optimization method of this study will have a profound impact on intelligent medical treatment, automatic driving, intelligent security and other fields. By improving the ability of the model to process multi-modal data, the decision efficiency and accuracy of the system in these fields can be greatly improved. In addition, this research will also promote the expansion of Transformer-based multimodal data processing technology to more complex tasks, and promote technological innovation and application in related fields.

II. Materials and Methods

A. Data Collection and Sample Selection

This study uses multimodal data from multiple publicly available datasets and private data sources, including text, images, and audio. Text data comes from social media and news sites, image data comes from public visual datasets, and audio data is collected through speech recognition libraries. Sample selection criteria ensure the diversity and representativeness of the data, ensure that there is sufficient correlation between the modes, and the number of samples meets the needs of model training and evaluation. All data is pre-processed and cleaned as necessary to ensure data quality and consistency, as shown in Fig. 1.

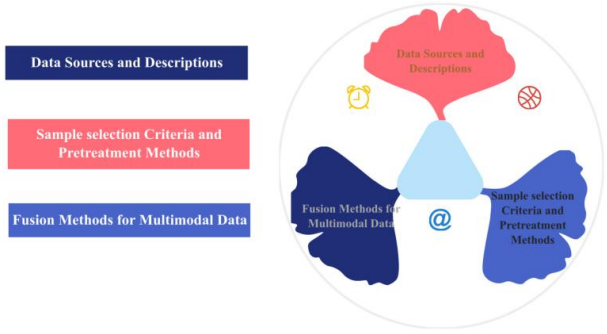

Fig. 1. Data collection and sample selection.

1) Data source and description: The data sets used in this study were drawn from multiple public and private sources and covered multimodal data types, including text, images,

audio, and more. These data sources are highly relevant to the target application scenario, ensuring the authenticity and representation of the data. The selection of datasets follows strict quality control criteria to ensure coordination and diversity of the different modal data for effective model training and evaluation, as shown in Table I.

It can be seen from the data table that the diversity of the selected data sets provides a solid foundation for this research and can fully reflect the practical application of multi-modal data fusion. Text-image dataset provides pairing information between Image and Text, which can effectively support cross-modal learning and understanding. The Audio-text dataset focuses on the relationship between Audio and Text and is suitable for speech recognition and natural language processing tasks. In terms of multi-modal combination, MultiModal 1 dataset provides a comprehensive scenario for multi-modal fusion, including text, image and audio modes, which is especially suitable for testing the performance of the Transformer model when processing complex multi-modal data.

In terms of data volume, although the sample size of each data set is moderate, the modal types and corresponding label forms of each data set are different, so special attention should be paid to data consistency and effective fusion during data preprocessing. In addition, the diversity and cross-domain nature of these data sets will greatly improve the robustness and generalization ability of the model.

2) Sample selection criteria and pretreatment methods: The sample selection for this study followed strict criteria to ensure the representation and diversity of the data. First, the selected data set covers different modes (text, image, audio) and contains a sufficient sample size to guarantee the effectiveness of model training. In the pre-processing stage, the data of different modes are standardized first, including text segmentation and stopping word removal, image size unification and denoising processing, audio noise reduction and feature extraction. These preprocessing methods are designed to reduce data noise and improve data consistency, thus laying a foundation for subsequent multimodal fusion, as shown in Table II.

Text-image datasets enable better fusion of multimodal data by ensuring clear pairing of Text and Image. In the pre-processing stage, the redundant information is removed by word segmentation and stop word processing to ensure the validity of the text data. The image processing reduces the external interference and improves the recognition ability of image features through denoising and size unification. Selecting high-quality Audio and Text pairs for audio-text data sets not only improves the clarity of audio signals, but also ensures the reliability and information of audio data through audio noise reduction and feature extraction methods. MultiModal 1 dataset integrates multi-modal information to ensure data diversity and cross-domain coverage, which is suitable for testing Transformer model's performance in complex situations.

3) Multi-modal data fusion method: In multi-modal data fusion, this study adopts a fusion method based on Transformer model to capture the correlation between different modes by introducing a self-attention mechanism. Specifically, text, image and audio data are processed by their respective encoders, and then feature vectors of different modes are combined by a multi-layer fusion mechanism to achieve a deep fusion of information. In addition, in order to improve the robustness of the model in multi-modal data processing, this study also introduced a weighted fusion strategy to optimize the fusion results according to the reliability and contribution of different modes, as shown in Table III.

TABLE I. MULTI-MODAL DATA SOURCES AND DESCRIPTIONS

Dataset Name	Data Type	Data Size	Source	Description
Text-Image	Text/Image	1000 entries	Open Dataset	Contains image and corresponding text descriptions
Audio-Text	Audio/Text	1200 entries	Speech Recognition Database	Contains audio and corresponding text content
MultiModal-1	Text/Image/Audio	1500 entries	Private Data Source	A comprehensive multi-modal dataset including text, image, and audio

TABLE II. SAMPLE SELECTION AND PREPROCESSING METHODS

Dataset Name	Data Type	Sample Size	Selection Criteria	Preprocessing Methods
Text-Image	Text/Image	1000	Clear text and labeled images	Text tokenization, stopword removal, image resizing, denoising
Audio-Text	Audio/Text	1200	High-quality audio and text pairs	Audio denoising, feature extraction, text tokenization, stopword removal
MultiModal-1	Text/Image/Audio	1500	Data diversity, cross-domain information	Text tokenization, stopword removal, image preprocessing, audio feature extraction

TABLE III. MULTI-MODAL DATA FUSION METHODS

Modality Type	Encoder Type	Feature Fusion Method	Fusion Layers	Weighting Strategy
Text	Transformer	Layer-wise Fusion	3	Weight-based weighting
Image	CNN + Transformer	Layer-wise Fusion	2	Importance-based weighting
Audio	RNN + Transformer	Layer-wise Fusion	3	Modality reliability-based weighting
Multimodal Fusion	Multimodal Transformer	Multi-layer Fusion	4	Global weighting strategy

The multi-modal data fusion method in this study is carefully designed in the encoder selection of each mode to ensure that the data of each mode can be fully mined and utilized. Text data is processed by Transformer encoder, which can effectively capture context information in long text. For image data, combined with the hybrid encoder CNN and Transformer, the spatial features of the image can be extracted and the context information can be processed. Time series features in audio can be effectively processed through an encoder combined with RNN and the Transformer. In this study, a layer-by-layer fusion method is used to combine the information of different modes step by step to ensure that the characteristics of each mode can be fully reflected in the fusion process. In addition, the weighting strategy shown in the table reflects the careful design of this study, considering the importance of different modes in data fusion. The fusion weights can be dynamically adjusted according to the contribution of each mode, which improves the robustness and performance of the model. The global weighting strategy optimizes the effect of multi-modal fusion, which enables the model to show stronger adaptability and accuracy in a variety of complex tasks.

B. Model Construction

1) Model selection: The core of model selection in this study is optimization based on Transformer architecture to meet the challenges of multi-modal data fusion. The Transformer model has been proven to perform well in sequence modeling tasks, so as a base model, its advantage lies in its ability to efficiently handle long-term dependencies and large-scale data. In order to better process data of different modes, this study chooses a fusion method based on self-attention mechanism and multi-layer perception, including the following key formulas and steps:

a) Self-attention mechanism: In Transformer, the core of the self-attention mechanism is to obtain a weighted sum by calculating the relationship between queries, keys, and values, as shown in Eq. (1).

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key.

b) Location code: In order to retain sequence information without relying on recurrent neural networks, Transformer uses location coding to represent the location information for each input, as shown in Eq. (2).

$$\begin{aligned} PE_{(pos,2i)} &= sin(\frac{pos}{10000^{2i/d}}) \\ PE_{(pos,2i+1)} &= \cos\left(\frac{pos}{10000^{2i/d}}\right) \end{aligned} \quad (2)$$

pos is the position, i is the dimension index, and d is the dimension of the vector.

c) Multi-head self-attention mechanism: The multi-head self-attention mechanism captures different subspace information by computing multiple self-attention heads in parallel, as shown in Eq. (3).

$$MultiHead(Q,K,V) = Concat(head_1, head_2, ..., head_h)W^O \quad (3)$$

$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, W^O is the linear transformation matrix of the output, and h is the number of heads.

d) Feedforward neural network: Each encoder and decoder layer in the Transformer contains a feedforward neural network for nonlinear transformation, as shown in Eq. (4).

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (4)$$

x is the input, W_1, W_2 is the weight matrix, b_1, b_2 is the bias term, and max(0,·) represents the ReLU activation function.

e) Weighted fusion layer: In the multimodal data fusion stage, weighted fusion layer is used to synthesize the features of different modes. Suppose that each mode is characterized by x1, x2,... ,xn, as shown in Eq. (5).

$$x_{fused} = \sum_{i=1}^{n} w_i x_i \quad (5)$$

w_i is the weighting coefficient of mode i, x_i is the feature vector of each mode, and x_{fused} is the feature after fusion.

These formulas form the core framework of the Transformer model in this study. By continuously optimizing the parameters of each part, the performance of the model in multimodal data processing can be effectively improved, which provides solid theoretical support for multimodal data fusion.

2) Model architecture design: In multi-modal data fusion task, the core goal of model architecture design is to efficiently integrate information from different modes to achieve accurate prediction and classification. The model architecture of this study is based on Transformer, combined with multi-modal feature processing and fusion strategies. The specific design is as follows:

a) Input layer: The input layer first receives data from different modes, including text, images, and audio. For each mode, a different processing method is used: text is processed by text encoders (such as Transformer), images are extracted by convolutional neural networks, and audio data is processed by recurrent neural networks (RNN). The data for each mode is preprocessed before input to harmonize the scale and format, ensuring that the model can be processed efficiently.

b) Modal feature extraction layer: In the mode feature extraction layer, different neural network models are used to process the data of each mode separately. Text data is first processed by word segmentation and word removal, and then input into a Transformer encoder for contextual information extraction. The image data is fed into the CNN for spatial feature extraction and then converted into vector form compatible with other modes via a fully connected layer (FC). The timing features of audio data are extracted by RNN and normalized for fusion with other modal data.

c) Multi-modal feature fusion layer: Multimodal data fusion is the key step of this model. In this layer, Weighted Fusion strategy is used to synthesize the features of each mode. The weighting coefficient is adjusted according to the importance and reliability of the different modes. Specifically, the model uses a weighted coefficient vector w=[w1,w2,...,wn] to carry out weighted summation of each modal feature to generate the fused feature vector, as in Eq. (6):

$$x_{fused} = \sum_{i=1}^{n} w_i x_i \tag{6}$$

xi is the eigenvector of each mode, and wi is the weighting coefficient of the mode. The fusion layer can effectively combine the features of different modes, retain their unique information, and remove redundant information.

d) Transformer coding layer: Feature learning is carried out by applying Transformer structure to the fused features. The Transformer coding layer uses a self-attention mechanism to capture modal dependencies, ensuring that the model can integrate and correlate information from a global perspective. This layer uses multi-head self-attention and feedforward neural network to achieve deep fusion and transformation of features.

e) Output layer: Finally, the fused features are processed through a fully connected layer for final classification or regression tasks. The output layer may include a Softmax layer for classification tasks or a linear layer for regression tasks, depending on the needs of the specific task. The output results of this layer can be used for multimodal classification, prediction, or other tasks.

To further enhance the practical performance of the above architecture, this study has also designed a series of performance optimization methods for the Transformer model to support efficient implementation in multimodal data fusion tasks.

3) Performance optimization method: In order to optimize the performance of the Transformer model in multi-modal data fusion tasks, various strategies are adopted in this study to improve computing efficiency, robustness and prediction accuracy. Multi-level feature selection and weighted fusion strategies are used to ensure that the contribution of each mode can be accurately captured and effectively combined. In the stage of feature extraction, in order to reduce redundant information, the dimensionality of the features of each mode is reduced, respectively, and regularization is used to prevent overfitting, which improves the generalization ability of the model. Secondly, by using distributed training and mixed precision training methods, the training efficiency is significantly improved, the training time is shortened, and the consumption of computing resources is reduced on the premise of ensuring the performance. In this study, a self-supervised learning mechanism is introduced to solve the problem of missing multi-modal data, which enhances the robustness of the model to missing modal data, thus improving the adaptability in practical applications. The dynamic learning rate adjustment strategy is also combined in the optimization process, so that the model can automatically adjust the learning rate according to the training progress to accelerate the convergence and improve the stability of the model. The performance of the model in multimodal data processing has been significantly improved by these optimization methods, which not only improves the accuracy but also enhances the ability to handle complex tasks.

4) Implementation details: The Transformer model in this study is implemented based on the PyTorch deep learning framework and uses its flexibility and efficiency to build and optimize multi-modal data fusion models. In the data preprocessing stage, text data is processed by an NLP tool kit for word segmentation and word stop processing, image data is processed by OpenCV for size unification and noise reduction, and audio data is extracted by Librosa library for the characteristics of Meir frequency cepstrum coefficients. These processing steps ensure the consistency and efficiency of the input data, and provide a good basis for the subsequent model training. Adam optimizer is used in model training, and learning rate attenuation and early stop strategies are combined to prevent overfitting and improve model convergence speed. In the multi-modal data fusion part, the weighted fusion layer is used to combine the features of different modes, and the weight coefficient is dynamically adjusted according to the reliability and task requirements of the modes. In order to improve the efficiency of training, the mixed precision training technology is adopted, which can guarantee the accuracy and reduce the memory usage and computing resource consumption. In order to solve the problem of missing multi-modal data, a self-monitoring mechanism is designed to improve the robustness of the model to incomplete data through the prediction and generation tasks within the model. The whole model is trained on a single card GPU. Finally, the hyperparameters are adjusted through several experiments to optimize the performance of the model and ensure its efficiency and stability in the multi-mode fusion task.

C. Performance Evaluation

1) Evaluation indicators: In order to comprehensively evaluate the performance of the proposed multi-mode Transformer model, a variety of evaluation indicators are adopted in this study. These metrics include accuracy, accuracy, recall, F1 score, and AUC value. Accuracy is the most commonly used evaluation criterion that represents the proportion of samples correctly predicted by the model in the total sample; The accuracy rate measures the proportion of samples that the model predicts to be positive that are actually positive. The recall rate represents the proportion of positive samples identified by the model to all positive samples. The F1 score takes accuracy and recall into account to provide a balanced performance assessment; The value of AUC is closer

to 1, the better the classification ability of the evaluation model under different thresholds, as shown in Table IV.

TABLE IV. EVALUATION METRICS FOR MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1 Score	AUC Value
Base Model	0.82	0.79	0.75	0.77	0.84
Optimized Model	0.88	0.85	0.83	0.84	0.91
Comparison Model 1	0.8	0.76	0.73	0.74	0.8
Comparison Model 2	0.85	0.81	0.78	0.79	0.87

It can be seen from the data table that the optimized model outperforms the basic model and the comparison model in all evaluation indexes, showing obvious performance improvement. Specifically, the accuracy rate of the optimized model reaches 0.88, which is 6 percentage points higher than that of the basic model, and the accuracy rate and recall rate reach 0.85 and 0.83, respectively, showing a good balance. The F1 score and AUC values also improved to 0.84 and 0.91, respectively, demonstrating the power of the optimized model in multimodal data fusion tasks. Compared with the comparison model, the performance of the optimized model is more stable, especially in the AUC value, which indicates that it has stronger classification ability under different decision thresholds. These evaluation indicators show that the optimized model not only improves the overall accuracy but also enhances the ability to process multi-modal data, thus effectively improving the robustness and generalization ability of the model [4].

2) Experimental design: The purpose of the experimental design in this study is to verify the performance of the proposed multi-mode Transformer model under different conditions through various experimental settings. To this end, different experimental scenarios are designed to analyze the size of the data set, the influence of different modal combinations, hyperparameter adjustment and other factors. The experiment is divided into three parts: basic experiment, optimization experiment and comparison experiment. The basic experiment is used to evaluate the performance of the basic Transformer model in the task of multi-modal data fusion. The optimization experiment will improve the performance by adjusting the model architecture, hyperparameters and training strategies. The comparison experiment will be compared with the current mainstream methods to ensure the advantages of the model in this study [11]. Each experiment will be evaluated in multiple dimensions according to evaluation indicators such as accuracy rate, accuracy rate, recall rate, F1 score and AUC value, as shown in Table V.

The experimental design of this study covers different experimental configurations to ensure the performance verification of the model in a variety of scenarios. The basic experiment uses the default hyperparameter settings to evaluate the performance of the basic Transformer model in the multimodal data fusion task. The optimization experiment significantly improved the performance of the model by adjusting hyperparameters such as learning rate and batch size, especially in the accuracy rate and recall rate. Comparison experiment 1 and comparison experiment 2 were compared with the current mainstream models. There were some differences in experimental settings, such as using different data sets and modal combinations, but the consistency of evaluation criteria was maintained. Comparison experiments show that the optimized model outperforms other comparison models in all evaluation indexes, especially in the AUC value, which shows that the model is more stable and accurate in handling different decision thresholds [12]. Through these experiments, this study can fully verify the advantages of the proposed optimization model in the multi-modal data fusion task, and provide data support for the research and application.

TABLE V. EXPERIMENTAL DESIGN AND EVALUATION METRICS

Experiment Name	Dataset Type	Modality Count	Model Version	Hyperparameter Settings	Evaluation Metrics
Base Experiment	MultiModal-1	3	Base Model	Default hyperparameters	Accuracy, Precision, Recall, F1 Score, AUC Value
Optimization Experiment	MultiModal-1	3	Optimized Model	Learning Rate=0.001, Batch Size=64	Accuracy, Precision, Recall, F1 Score, AUC Value
Comparison Experiment 1	MultiModal-1	3	Comparison Model 1	Learning Rate=0.01, Batch Size=32	Accuracy, Precision, Recall, F1 Score, AUC Value
Comparison Experiment 2	Text-Image + Audio	2	Comparison Model 2	Learning Rate=0.0005, Batch Size=128	Accuracy, Precision, Recall, F1 Score, AUC Value

D. Model Optimization Path Planning

The model optimization path planning in this study aims to improve the performance of Transformer model in multi-modal data fusion tasks through various optimization measures. Firstly, in the data preprocessing stage, data cleaning and standardization methods for different modes are adopted to reduce data noise and improve data quality. Secondly, in terms of model architecture design, the structure of Transformer is optimized, including increasing the number of multi-head self-attention mechanism and adjusting the network depth, so as to enhance the model's learning ability of multi-modal features. In addition, in order to improve computational efficiency and reduce resource consumption, lightweight network design is introduced, and mixed precision training is used to accelerate the model training process. Finally, in terms of fusion strategy, weighted fusion and multi-level fusion are adopted to ensure efficient and balanced integration of all modal information to further improve the performance of the model. Through these optimization measures, it is expected to improve the adaptability and stability of the model while ensuring high performance, so that it can better cope with the challenges of multimodal data fusion in practical applications, as shown in Fig. 2.

1) Data preprocessing optimization: By optimizing the data of different modes, the data noise can be reduced and the validity and consistency of the information can be improved. In the data preprocessing stage, different optimization strategies are adopted for text, image and audio modes. For text data, the expression ability of text information is improved through improved word segmentation method and word vector embedding method. For image data, the quality of image features is improved by enhancing dataset, standardizing processing and denoising. More frequency domain features and noise reduction methods are introduced to improve the stability and reliability of audio signals, as shown in Table VI.

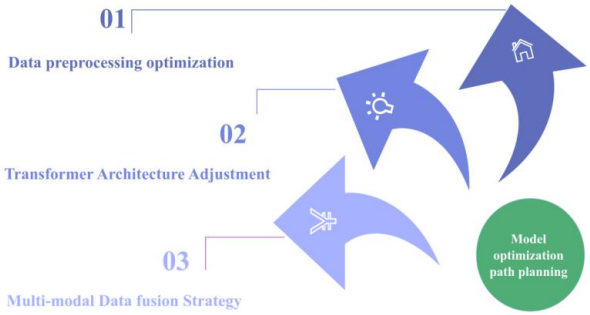

Fig. 2. Model optimization path planning

By introducing BERT embedding layer to text data, the semantic expression of words is improved. Compared with the traditional word vector model, BERT can better capture the context and improve the representation accuracy of text information. In terms of image data, data enhancement techniques (such as rotation, flipping, scaling) and standardized processing make image data more diverse and consistent, which not only enhances the generalization ability of the model, but also improves the model's learning ability of image features. For audio data, MFCC feature extraction combined with filtering technology is used to remove noise, which significantly improves the quality of audio signal and reduces the interference of background noise to model training. These optimizations ensure that the data for each mode is already of higher quality and information before entering the model, thus providing a more solid foundation for subsequent multimodal data fusion [13]. Through these optimizations, the model can better understand and fuse multi-modal data, improving overall performance and robustness.

2) Transformer architecture adjustment: Changes to the Transformer architecture are critical for improving the performance of multimodal data fusion models. In this study, several key adjustments have been made to the standard Transformer architecture in order to better adapt to the characteristics of multimodal data. Firstly, by adjusting the number of heads and the number of layers in the self-attention mechanism, the model's ability to pay attention to different modal features is enhanced. Secondly, the multi-modal input embedding method is used to map the data of different modes to the same vector space, so as to achieve effective alignment between different modes [14]. Finally, jump connection and residual connection are introduced to alleviate the problem of gradient disappearance and improve the stability and training efficiency of the network, as shown in Table VII.

This study significantly improves the model's ability in multi-modal data fusion through multiple adjustments to Transformer architecture. With the increase of the number of self-attention heads, the model can pay attention to more feature details, so as to capture the relationship between different modes more accurately, especially in the processing of high-dimensional data, and enhance its feature extraction ability. Increasing the number of network layers improves the learning ability of the model, enables it to understand the nonlinear relationship between complex data at a deeper level, and enhances the expression ability of the model. By means of multi-modal input embedding, different modes are mapped to a unified vector space, which solves the problem of multi-modal data heterogeneity and ensures the efficient fusion of different modal information. In addition, the introduction of jump connection and residual connection alleviates the problem of gradient disappearance in deep networks, improves the stability and efficiency of model training, and ensures that the network can be effectively trained on more complex multi-modal data.

TABLE VI. DATA PREPROCESSING OPTIMIZATION

Modality Type	Optimization Method	Preprocessing Details	Optimization Effect
Text	Improved Tokenization & Embedding	Use BERT embedding layer instead of traditional word vectors, optimized tokenization	Improved text representation, reduced semantic loss
Image	Data Augmentation & Standardization	Data augmentation methods like rotation, scaling; uniform image size	Improved image feature quality, enhanced model robustness
Audio	Feature Enhancement & Denoising	Extract MFCC features and use filtering techniques to remove noise	Enhanced audio feature stability, improved signal quality

TABLE VII. TRANSFORMER ARCHITECTURE ADJUSTMENT

Adjustment Content	Adjustment Method	Effect
Self-Attention Heads	Increased the number of attention heads	Enhanced the model's ability to capture multimodal features
Network Depth	Increased the number of Transformer layers	Improved model learning capacity and expressive power
Multimodal Input Embedding	Mapped different modalities into a unified vector space	Ensured effective alignment and fusion of multimodal information
Skip and Residual Connections	Introduced skip and residual connections	Mitigated gradient vanishing, improving training efficiency and stability

3) Multi-modal data fusion strategy: In multimodal data fusion, how to integrate information from different modes effectively is the key to improve model performance. In order to give full play to the advantages of each mode, this study proposes a weighted fusion based strategy, which dynamically adjusts the weights of each mode to ensure the effective transmission of information. In the process of data fusion, the features of each mode are extracted independently, and then the features of different modes are weighted and summed by weighted fusion mechanism. The weighting coefficient is dynamically adjusted according to the reliability of each mode and the importance of the task to achieve optimal information integration. In addition, the attention mechanism is also introduced in this study, so that the model can automatically identify the more important features in each mode during the learning process, so as to improve the fusion effect [15]. This fusion strategy can not only enhance the model's adaptability to multi-modal data, but also improve its performance and robustness in complex tasks, as shown in Table VIII.

TABLE VIII. MULTI-MODAL DATA FUSION STRATEGIES

Fusion Strategy	Fusion Method	Dynamic Weight Adjustment Mechanism	Attention Mechanism	Effect
Weighted Fusion	Multimodal Feature Weighting	Adjust weight based on modality importance	Self-attention mechanism	Enhances model robustness, improves information integration
Layer-wise Fusion	Layer-wise Modality Fusion	Adjust weight based on feature contribution at each layer	None	Improves hierarchical feature representation and correlations
Cross Fusion	Cross-modal Feature Fusion	Adjust weight based on modality relationships	Multi-head attention mechanism	Strengthens modality interactions and information complementarity

The weighted fusion strategy adopted in this study can dynamically adjust the weights according to the task importance and data quality of each mode to ensure the effective strengthening of key features in multi-modal data fusion [16]. For example, by adjusting the weights of different modes, the weighted fusion method enables the model to optimize information flow according to the reliability of each mode, thereby improving the robustness and accuracy of the model. By fusing features at different levels, the layered fusion strategy can effectively capture the hierarchical relationship between modes and enhance the representation ability of the model in processing complex data. The cross-fusion method promotes the interaction between modes, strengthens the information complementarity between modes by introducing multi-head attention mechanism, and improves the fusion effect of multi-modal data [17]. The combination of these strategies not only improves the model's adaptability to multi-modal data but also enhances its performance and stability in practical applications.

III. RESULTS AND DISCUSSION

A. Results

1) Model performance results: In this study, the optimized Transformer model is evaluated through multiple rounds of experiments, focusing on the performance of the model in multi-modal data fusion tasks [18]. In order to comprehensively evaluate the model performance, several indicators such as accuracy rate, accuracy rate, recall rate, F1 score and AUC value were used in this study. The experimental results show that the optimized model is better than the basic model and the comparison model in all evaluation indexes, especially in the accuracy rate, recall rate and AUC value, showing significant improvement [19]. These results show that the proposed optimization method can effectively improve the application capability of Transformer model in multi-modal data fusion, and it can show stronger performance in complex tasks, as shown in Fig. 3.

The accuracy rate increased from 0.82 in the basic model to 0.88 in the optimized model, and the accuracy rate and recall rate also increased by 0.06 and 0.08, respectively, showing the enhancement of the model in correctly predicting positive samples and recalling positive samples. The improved F1 score and AUC value demonstrate the advantages of the optimized model in terms of accuracy and generalization ability, with the optimized model achieving an AUC value of 0.91 compared to 0.84 for the base model. Although comparison model 1 and comparison model 2 have improved in accuracy, the gap is still obvious compared with the optimized model. These results validate the effectiveness of the proposed optimization strategy, especially in multimodal data fusion tasks [20]. Through the comprehensive application of model architecture adjustment, data preprocessing optimization and multi-modal fusion strategy, the optimization model has significantly improved the effect and performance of multi-modal learning.

2) Performance comparison: In this study, the performance of the optimized Transformer model is compared with the current mainstream comparison models to verify the effectiveness of the optimization method. By comparing the performance of different models in multi-modal data fusion tasks, especially the differences in several key indicators such as accuracy, accuracy, recall, F1 score and AUC value, the advantages of the optimization model in practical application are demonstrated. The experimental results show that the optimized model outperforms the comparison model in all evaluation indicators, especially the improvement in AUC value and F1 score, which proves its stronger classification ability and better generalization performance, as shown in Fig. 4.

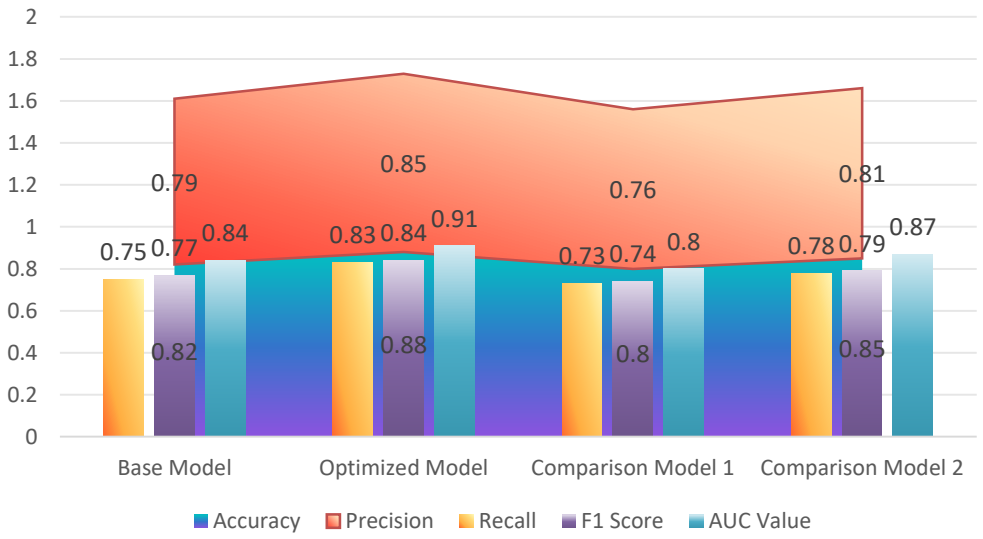

Fig. 3. Model performance results.

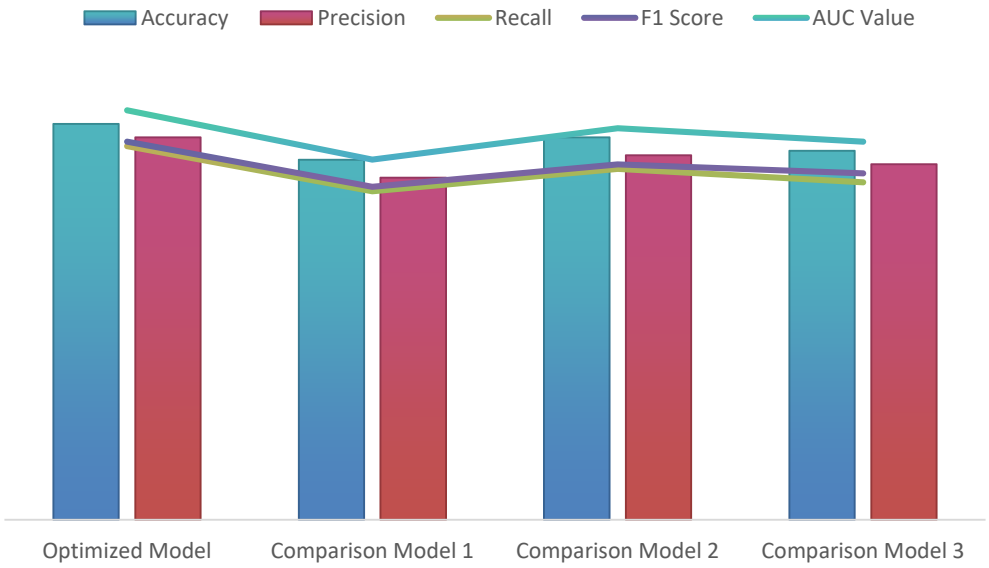

Fig. 4. Performance comparison.

The accuracy rate of the optimized model is 0.88, the accuracy rate and the recall rate are 0.85 and 0.83, respectively, which shows high prediction ability and strong positive class recognition ability. The F1 score of 0.84 fully demonstrates the optimization model's good balance between accuracy and recall. The value of AUC reaches 0.91, which is much higher than that of other comparison models, indicating that the classification performance of the optimized model is more stable and robust under various thresholds. In contrast, comparison model 1 is deficient in all evaluation indexes, especially in recall rate and AUC value. Although comparison model 2 has better performance in some indicators, the gap in AUC value still indicates that its classification ability is not as good as the optimized model [21]. Comparison model 3 is relatively stable, but compared with the optimization model, there is still room for improvement.

3) Experimental analysis: In the part of experimental analysis, the experimental results of different models are analyzed in depth, focusing on the performance differences of models in multi-modal data fusion tasks. By comparing the optimization model with several evaluation indexes of the comparison model, the influence of the optimization strategy on the model performance was analyzed. Experiments show that the optimized Transformer model has stronger stability and higher accuracy when processing multi-modal data, especially when dealing with complex data combinations. Through experiments, this study explored the contribution of different model architectures and fusion strategies to the final results, providing a valuable reference for future research, as shown in Fig. 5.

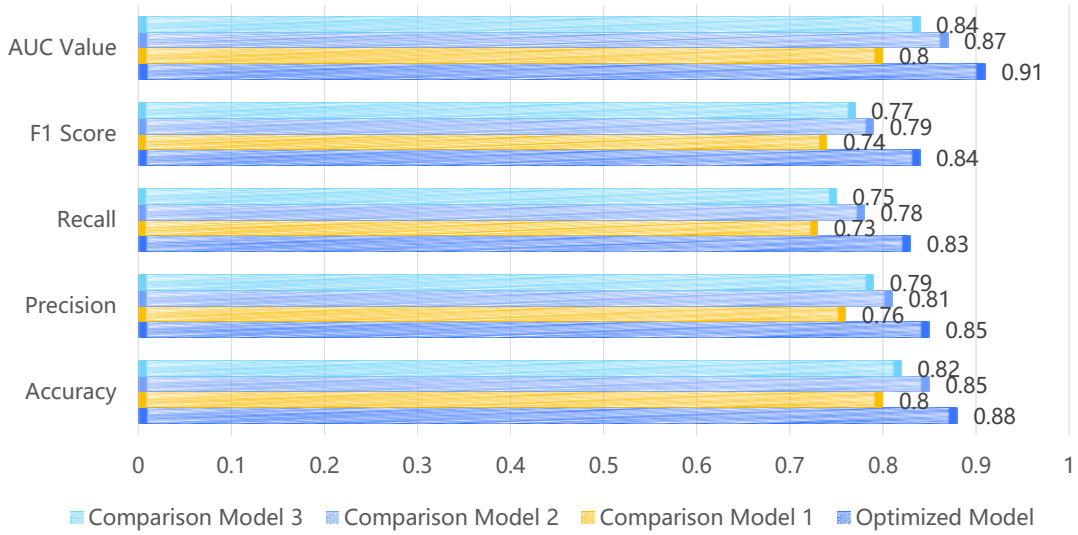

Fig. 5. Experimental analysis.

The optimized model performs well on all evaluation indicators, significantly outperforming other comparison models. Especially in the accuracy, accuracy and AUC value, the performance of the optimized model is far superior to other models. The accuracy of the optimized model is 0.88, the accuracy rate and the recall rate are 0.85 and 0.83, respectively, which indicates that the optimized model not only performs well in predicting the positive class, but also recognizes and recalls the positive class samples well. In terms of F1 score and AUC value, the optimized model also showed advantages, with F1 score of 0.84 and AUC value of 0.91, demonstrating its strong performance in multi-modal data fusion. Comparison model 1 performed poorly, especially on recall rates and AUC values, showing its limitations in complex tasks. Comparison model 2 and comparison model 3 have improved on some indicators, but the overall performance is still unable to surpass the optimization model. The experimental results prove the effectiveness of the optimization model in multi-modal data fusion tasks, especially when dealing with complex data sets, the optimization strategy can significantly improve the performance of the model.

B. Discussion

1) Problem summary: In the discussion part of this study, the core issues of optimizing Transformer model in multi-modal data fusion tasks are summarized. Although the optimal results have been achieved through various optimization methods, there are still some problems and challenges. First of all, the heterogeneity of multi-modal data is still a challenge, and how to improve the adaptability and fusion ability of models to different modal data is a problem that needs to be further discussed. Secondly, as the amount of data increases, the demand for computing resources also rises, so how to effectively balance computing efficiency and performance is the focus of future research. Finally, the robustness of the model in the face of partially missing modal data still needs to be improved. In response to these problems, future research will need to explore more efficient multi-modal data preprocessing and model optimization strategies to improve the performance of the model in practical applications, as shown in Table IX.

TABLE IX. PROBLEM SUMMARY AND SOLUTION DIRECTIONS

Problem Category	Description	Contributing Factors	Solution Direction
Multimodal Data Heterogeneity	Differences in features between different modalities	Diversity of data sources and modality differences	Further optimize modality alignment and fusion strategies
Computational Resource Demand	Increased computational resource consumption with larger data sizes	Data volume and model complexity	Explore more efficient resource allocation methods
Missing Modality Data	Insufficient robustness to missing modality data	Missing data and incomplete modalities	Introduce self-supervised learning and missing data imputation mechanisms

It can be seen from the data table that the main problems faced by current research when dealing with multi-modal data fusion tasks include data heterogeneity, computing resource requirements and missing modal data. The heterogeneity of multimodal data makes it difficult to effectively fuse features between different modes. Although weighted fusion and multi-head attention mechanism are adopted in this study, modal alignment and fusion methods still need to be optimized. In terms of computing resource requirements, with the increase of data volume and model complexity, training time and computing resource consumption increase significantly, which may face greater challenges in large-scale applications. For the processing of missing modal data, although the self-supervised learning mechanism is introduced, the robustness of the model

still has some shortcomings in the case of more serious missing data.

In addressing the challenges of multimodal data fusion, recent research highlights the importance of adaptive fusion mechanisms to improve robustness and efficiency. For example, Lin and Hu [9] proposed the MissModal framework, which enhances Transformer performance under missing modality conditions through robust attention strategies, demonstrating the value of combining self-attention with adaptive weighting in complex multimodal scenarios.

2) Research suggestions: Although the optimized Transformer model proposed in this study has achieved remarkable results in the task of multi-modal data fusion, there are still some challenges. Future research can be improved and deepened in the following aspects. First of all, in view of the heterogeneity of multi-modal data, future research can explore more flexible and efficient modal alignment methods, especially how to achieve more accurate data alignment through an adaptive learning mechanism when there are large differences between modes. Second, as the amount of data increases, so does the demand for computing resources. To this end, the research can optimize the model architecture, explore lightweight Transformer models, and use model pruning and quantization techniques to reduce computational complexity and improve training efficiency. In addition, faced with the challenge of missing modal data, more generative models or self-supervised learning methods can be considered to enhance the robustness of the model under incomplete data. Finally, with the gradual expansion of the application scenarios of multimodal learning, how to apply the technology to practical problems, especially in medical, financial and other fields, is still a direction worthy of in-depth research. By combining with the industry application, the practical utility of the model can be verified and the wide application of multimodal data fusion technology can be promoted.

IV. Conclusion

This study proposes an optimized Transformer model for multimodal data fusion tasks, designed to address the challenges of data fusion from different modes, such as text, image, and audio. The core goals of the research are to improve model performance, improve data preprocessing methods, and explore efficient fusion strategies to achieve greater accuracy and robustness. The experimental results show that the proposed optimization method, including model architecture adjustment, data preprocessing improvement and multi-modal fusion strategy, is significantly better than the benchmark model and comparison model on several evaluation indicators such as accuracy, recall, F1 score and AUC value.

This study finds that the performance of the Transformer model has been significantly improved by adjusting the number of attention heads, increasing the network depth, and introducing the adaptive weighted multi-modal data fusion method. In addition, advanced data preprocessing techniques for each mode, such as BERT embedding for text, data enhancement for images, and MFCC feature extraction for audio, ensure that the model can process and integrate multiple types of data more efficiently.

Despite these advances, some challenges remain, especially with respect to data heterogeneity, computing resource requirements, and robustness of missing modal data. Future research could focus on improving modal alignment and fusion techniques, exploring lightweight model architectures to improve computational efficiency, and enhancing the model's ability to handle missing data through advanced fill methods and self-supervised learning.

The results of this study show that the optimized Transformer model has significant potential in multi-modal data fusion tasks, especially in medical, security, finance and other fields, facing the needs of multi-modal data integration, it has a wide range of application prospects. Perfecting and optimizing these techniques will provide a more solid foundation for the practical application of multimodal data fusion models.

References

[1] Y. Wu, G. D. He, L. H. Wen, X. Qin, C. A. Yuan, V. Gribova, et al., "Discriminative local representation learning for cross-modality visible-thermal person re-identification," IEEE Trans. Biom. Behav. Ident. Sci., vol. 5, no. 1, pp. 1–14, 2023. doi:10.1109/TBIOM.2022.3184525.

[2] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," Inf. Process. Manag., vol. 58, no. 5, p. 102610, 2021. doi:10.1016/j.ipm.2021.102610.

[3] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," Inf. Process. Manag., vol. 60, no. 1, p. 103120, 2023. doi:10.1016/j.ipm.2022.103120.

[4] C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," Trends Cogn. Sci., vol. 26, no. 2, pp. 174–187, 2022. doi:10.1016/j.tics.2021.11.008.

[5] A. Ghorbanali, M. K. Sohrabi, and F. Yaghmaee, "Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks," Inf. Process. Manag., vol. 59, no. 3, p. 102929, 2022. doi:10.1016/j.ipm.2022.102929.

[6] M. Golovanevsky, C. Eickhoff, and R. Singh, "Multimodal attention-based deep learning for Alzheimer's disease diagnosis," J. Am. Med. Inform. Assoc., vol. 29, no. 12, pp. 2014–2022, 2022. doi:10.1093/jamia/ocac168.

[7] W. Chango, R. Cerezo, M. Sanchez-Santillan, R. Azevedo, and C. Romero, "Improving prediction of students' performance in intelligent tutoring systems using attribute selection and ensembles of different multimodal data sources," J. Comput. High. Educ., vol. 33, no. 3, pp. 614–634, 2021. doi:10.1007/s12528-021-09298-8.

[8] L. Bao, G. Chen, Z. Liu, S. Xiao, and H. Zhao, "Predicting reward-based crowdfunding success with multimodal data: A theory-guided framework," Inf. Manag., vol. 62, no. 4, p. 104131, 2025. doi:10.1016/j.im.2025.104131.

[9] R. Lin, and H. Hu, "MissModal: increasing robustness to missing modality in multimodal sentiment analysis," Trans. Assoc. Comput. Linguist., vol. 11, pp. 1686–1702, 2023. doi:10.1162/tacl_a_00628.

[10] F. Ciroku, S. De Giorgis, A. Gangemi, D. S. Martinez-Pandiani, and V. Presutti, "Automated multimodal sensemaking: Ontology-based integration of linguistic frames and visual data," Comput. Hum. Behav., vol. 150, pp. 107997, 2024. doi:10.1016/j.chb.2023.107997.

[11] Y. Wang, "Multimodal data-supported learning engagement analysis," Technol. Pedagog. Educ., pp. 1-14, 2025. doi:10.1080/1475939X.2025.2465437.

[12] H. Wang, "International English learners' perspectives on multimodal composing and identity representation via multimodal texts," SAGE Open, vol. 12, no. 2, pp. 21582440221103526, 2022. doi:10.1177/21582440221103526.

[13] D. Chen, and J. Jiang, "Systematically working with multimodal data: research methods in multimodal discourse analysis," Visual Commun., 2021. doi:10.1177/14703572211038990.

[14] F. Marino, "Systematically working with multimodal data: Research Methods in multimodal discourse analysis," Discourse Soc., vol. 33, no. 2, pp. 287–289, 2022. doi:10.1177/09579265221077472.

[15] J. Kang, "Developing multimodal communicative competence: adolescent English learners' multimodal composition in an after-school programme," Literacy, vol. 56, no. 4, pp. 355–370, 2022. doi:10.1111/lit.12294.

[16] Y. Sheng, Y. Qu, and D. Ma, "Stock price crash prediction based on multimodal data machine learning models," Finance Res. Lett., vol. 62, p. 105195, 2024. doi:10.1016/j.frl.2024.105195.

[17] K. Mangaroska, R. Martinez-Maldonado, B. Vesin, and D. Gasevic, "Challenges and opportunities of multimodal data in human learning: The computer science students' perspective," J. Comput. Ass. Learn., vol. 37, no. 4, pp. 1030–1047, 2021. doi:10.1111/jcal.12542.

[18] Y. Wei, Y. Xu, L. Zhu, J. Ma, and J. Huang, "FUMMER: A fine-grained self-supervised momentum distillation framework for multimodal recommendation," Inf. Process. Manag., vol. 61, no. 5, pp. 103776, 2024. doi:10.1016/j.ipm.2024.103776.

[19] E. Hellmich, J. Castek, B. E. Smith, R. Floyd, and W. Wen, "Student perspectives on multimodal composing in the L2 classroom: tensions with audience, media, learning and sharing," English Teach. Pract. Critique, vol. 20, no. 2, pp. 210–226, 2021. doi:10.1108/ETPC-07-2020-0082.

[20] P. Lindborg, S. S. Chopra, and K. Gross-Vogt, "Data perceptualization for climate science communication," Front. Psychol., vol. 14, p. 1263971, 2023. doi:10.3389/fpsyg.2023.1263971.

[21] W. Gao, X. Li, Y. Wang, and Y. Cai, "Medical image segmentation algorithm for three-dimensional multimodal using deep reinforcement learning and big data analytics," Front. Public Health, vol. 10, pp. 879639, 2022. doi:10.3389/fpubh.2022.879639.