In the field of human-computer interaction, human intention detection is a challenging problem and a key link to achieve barrier-free communication between humans and machines. With the rapid evolution of multimedia technology and the widespread use of social media platforms, the detection of user intent has become increasingly challenging. Traditional unimodal approaches, especially those relying solely on either textual or visual information, often fall short in capturing the intricacies of user intentions in multimedia content. To address this limitation, the fusion of image and text modalities using multimodal technology has emerged as a promising solution for intent detection.

Compared with single-modal data such as images and text, multi-modal data can contain more information and can more accurately identify user intentions. At present, there are few studies on intent recognition based on image and text fusion, and they mainly consider how to integrate modal features in the feature fusion stage, while in the feature extraction stage, only a single-modality pre-trained model is used, which not only increases the complexity of the system , which increases the computational cost and may also limit the model’s comprehensive understanding of the global context, making it difficult to effectively capture the correlation information between modalities.

This research endeavors to design a new intention detection framework decomposing the multimodal learning problem into two equally important stages of representation and fusion, to explore the integration of image and text data to enhance the accuracy and robustness of intent detection in multimedia content. In the feature representation part, the CLIP multi-modal large-scale pre-training model is used to simultaneously extract text and image features, which simplifies system integration and saves computing resources while learning the associated information between modalities. In the feature fusion part, due to the different importance of text and pictures, an attention-based cross-modal fusion method is designed, which enables the model to dynamically adjust the attention to different modalities at each moment and capture important features, while reducing noise. In order to verify the effectiveness of the proposed model, this paper conducts experimental verification on the intent detection data set.

As an important research direction in the fields of artificial intelligence and computer science, intent recognition aims to enable computer systems to understand and interpret users' true intentions, thereby responding to user needs more intelligently. In recent years, with the rapid development of multimedia technology, the forms of information released by users have become more diverse. When many users publish text information, they usually add corresponding picture information to express their true intentions more vividly and intuitively. This kind of Information in the form of multimedia better meets the needs of users to express themselves, obtain information and participate in interactions on social media. It also brings new challenges and opportunities to intent recognition.

In recent years, machine learning has made remarkable progress in processing various forms of media such as images and texts. Especially, the wide application of deep learning technology provides a powerful tool for intention detection, so that the model can better learn and understand the real purpose of users from complex massive data. The application of these technologies promotes the progress of intent detection technology in various fields, and provides more accurate and intelligent user interaction experience for intelligent systems. However, traditional intention recognition methods are usually limited to single-modal data analysis, and do not make full use of the rich information of multimedia data. In practical applications, users often use multiple forms of communication such as images and texts at the same time, and single-modal processing methods are difficult to fully understand and recognize the user's real intention.

In the era of digital multimedia, the field of intention recognition is facing more complex and diverse user expressions. As a cutting-edge research method, multi-modal data fusion, especially the fusion of image and text, provides a new idea to solve this problem. Compared with a single modality, image-text fusion contains more information. On the one hand, this information can complement each other. When the intention of a modality is unclear, it can be supplemented by other modalities. On the other hand, there may be false data or contradictory phenomena. How to correctly use and deal with the relationship and contradiction between each modality to improve the accuracy and robustness of intention recognition has become one of the main research directions in this field.

Multimedia includes text, image, audio, video, and other forms of information expression. In recent years, the explosive growth of social media platforms has provided a wide range of dissemination platforms for multimedia content, while multimedia forms of information better meet the needs of users to express themselves, obtain information, and engage in interaction on social media, but also bring challenges, especially in understanding and predicting user intentions in these digital Spaces. As one of the core technologies of human-computer interaction, intent detection aims to accurately identify the user's intent from his input to achieve more intelligent services. In multimedia technology, understanding whether a user's comment is truly informative, ironic, supportive, or critical through intent detection is crucial for improving recommendation systems, content moderation and fostering healthier online communities. However, there are some potential weaknesses in traditional intent recognition methods, such as only single modal data such as text or image are often considered, and the advantages of multimedia data are not fully utilized. Existing research based on image-text information fusion mainly focuses on sentiment analysis and content classification, but there are still significant gaps in interpreting the subtle intentions behind user behavior. Based on this, this paper proposes an intention recognition method based on image and text fusion. Through this research, we can not only verify the feasibility and effectiveness of image and text information fusion technology in intention recognition, improve the accuracy of intention recognition in multimedia environment, but also make it more suitable for diverse user communication scenarios in the real world.

The traditional intent recognition method is mainly based on manual rules and infers the user's intention through a pre-defined rule set. Although this method is simple, it requires manual arrangement of the rule set, has limited coverage and is not flexible enough. With the development of deep learning technology, researchers have proposed many intent recognition methods based on deep learning [13, 14]. These methods can automatically learn relevant patterns from massive data and are more efficient and accurate than manual rules. However, using only a single modality (for example, text modality) for intent recognition often cannot fully utilize the diversity of information, and it is difficult to effectively complete the task in the face of the diversity and complexity of input content. Therefore, how to effectively integrate and utilize multi-modal data in a multimedia environment is one of the main research issues in this field.

Nowadays，various multi-modal technology related solutions and methodologies have emerged to address the above problem in intent detection, for instance:

In [1], a model is proposed to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. While this model integrated text and images information to identify the intent, in the multimodal fusion stage, only the simple fusion strategy is used, that is, adds the two vectors, and the interaction information between text and images is ignored.

In [2] introduces a late-fusion approach for integration of the video signal with the captions signal for Intent Detection. Although it shows significant improvements with unimodal pretrained models, the HERO used in the article is a pretrained model only for video language, heterogeneity between modalities is not considered and its performance on image and text data is not yet known.

In [3] develops an adaptive multimodal fusion method based on an attention-based gated neural network, which can distinguish the contributions of different modalities. It designs complex strategies in feature fusion to reduce possible noise but use original pretrained models Bert and ResNet50 to extract text and image features respectively in feature extraction, which limits the model's comprehensive understanding of the global context.

In response to these challenges of insufficient single-modal detection capability and the limitations evident in the aforementioned multimodal methods, this research endeavors to design a new intention detection framework decomposing the multimodal learning problem into two equally important stages of representation and fusion. In the feature representation stage, we use a multimodal large-scale pre-trained model to extract features and achieve multimodal representations, and in the fusion stage, since the importance of text and pictures is not the same, we design an cross modality fusion method based on attention, this research aims to bolster the overall effect of intention detection.

**Research Questions**

1. How can the multimodal large-scale Pretrained models be leveraged for feature extraction and multimodal representation in the field of intention detection.
2. How to develop the proposed intention detection framework?
3. How can the proposed framework affect the accuracy of intention detection?

**Research Objectives**

1. To introduce and fine-tune multi-modal large-scale pre-training models to extract text and image features to achieve multi-modal representation.
2. To develop the proposed intention detection framework based on image and text fusion.
3. To evaluate the performance of the proposed intention detection framework by comparing its accuracy with the baseline model.

Recently, with the rapid development and application of multimedia technology, users are now more inclined to express their intentions through multimodal data such as text and images. In fact, multimodal data contains richer information, and the accuracy of intention detection can be improved by learning from multimodal data. At present, intention detection based on image and text fusion has become a research hotspot in the field of artificial intelligence. In the early days, researchers used machine learning methods to detect intent. For example, 【1】 compared the effects of bag-of-words model, TF-IDF and n-gram methods in short text intent analysis.【2】employ continuous bag-of-words coupled with support vector machines (SVM) to tackle the problem of intent classification.

With the development of deep learning, many intention recognition methods based on deep learning have been proposed 【3】, such as, 【4】 present a novel intent detection system which is based on a self-attention network and a Bi-LSTM，【5】propose a novel approach to intent recognition which involves combining transformer architecture with capsule networks。【6】developed an Intent Classification Model using BERT for the classification of Questions received from the Users or Humans to specific intents regarding the usage of specific features and components of the car。【7】 introduce intent detection methods backed by pretrained dual sentence encoders such as USE and ConveRT。

In recent years, multimodal technology has developed rapidly and become a research hotspot in the field of artificial intelligence. It has been widely applied in multiple fields. For example, in emotion recognition【8】, multimodal technology can be used to analyze text and image information, identify users' emotional tendencies and expressions. In terms of humor detection【9】, various information such as text, speech, and facial expressions are used to determine whether a sentence or situation is humorous. However, few studies have applied multimodal techniques to intention recognition. 【10】proposed a model to capture the complex meaning multiplication relationship between image and text in multimodal Instagram posts. 【11】proposed a late-fusion approach for integration of the video signal with the captions signal for Intent Detection.【12】introduced an adaptive multimodal fusion method based on an attention-based gated neural network, which can distinguish the contributions of different modalities.

With the gradual maturity of pre-training model technology in the field of natural language, multi-modal pre-training models have gradually attracted attention, and a series of visual-language pre-training work has emerged. Vision-and-Language Pre-training VLP (Vision-and-Language Pre-training) [3] refers to a universal representation of cross-modal training based on massive image-text data. The resulting pre-training model can be directly fine-tuned to adapt to downstream vision- language tasks. According to the different encoding methods, it can be roughly divided into twin-tower encoding and fusion encoding.

Twin-tower coding mainly focuses on the representation alignment of the respective modal encoding of images and texts, using the simplest dot product fusion features. Currently hot models such as CLIP [2] and ALIGN [4], etc., this type of method uses contrastive learning for pre-training, uses cosine similarity to measure the distance between modalities, and has demonstrated excellent performance in different fields. Recently, Meta AI He Kaiming's team launched the FLIP [24] multi-modal pre-training model, which integrates the image-text double masking technology in MAE [25] and can learn from more image-text data sets in a limited time, and effectively improves the efficiency of model pre-training compared with CLIP.

The fusion coding framework uses the Transformer mechanism for cross-modal fusion. ViLBERT [27] and LXMERT [28] proposed to use three different Transformers for image coding, text coding and feature fusion respectively. After increasing the network depth in the fusion stage, the hybrid coding model framework performed well in visual-language downstream tasks, shows excellent characterization capabilities. However, this type of algorithm is limited by network training and inference speed, and has not been widely used in the industry. ViLT [32] is optimized for the inference speed problem. Through a simplified network design, the encoder of the Transformer model is used to extract and process visual features instead of a separate computer vision model to extract features. Experiments show that this method can significantly reduce the number of parameters and running time, and the model effect is significantly better than fusion coding frameworks such as LXMERT, but there is still a certain gap between it and the CLIP [2] twin-tower framework.

After analyzing representative methods in the field of intent detection in recent years, it can be concluded that the research content of multimodal technology mainly includes three parts: feature extraction, multimodal representation, and multimodal fusion. However, most researchers only focus on the multimodal fusion part and propose some new methods, but the multimodal feature extraction and representation part is less considered and only the traditional feature extraction method is used. However, the development of multi-modal large-scale pre-trained models gives us new ideas for feature extraction and representation.

This research aims at detecting intent using image and text multimodal data. The main architecture is shown in Figure 1, which mainly includes three parts: feature extraction, multi-modal fusion, and classification. First, in the first part, text and image feature extraction, alignment, and multimodal representation are automatically achieved using the CLIP multimodal large-scale pre-trained model. Secondly, in the second part, considering that different modalities are data of different natures, contain different amounts of information, and have different contributions to intent detection, multi-modal feature fusion method based on cross-modal attention mechanism is designed. Finally, the fused features are input into the classifier to achieve intent recognition.

The quality of input features has an important impact on the prediction results of multi-modal intent recognition models. As early as the machine learning period, feature engineering determined the upper limit of learning. Better features mean you don’t need complex models to get excellent results. With the development of deep learning neural networks, the method of feature extraction has also changed greatly. Currently, in multi-modal intent recognition, BERT and ResNet pre-trained models are mainly used to extract text and image features. BERT and ResNet are usually trained independently, this results in each model only understanding the information of its specific modality, which limits the comprehensive understanding of the global context of the model, and the association information between modalities is very critical in image-text tasks. The multi-modal pre-training model has some obvious advantages over the single-modal pre-training model. It uses contrastive learning and other methods to learn the correlation information between modalities in the pre-training stage, so it can process multi-modal data at the same time and improve the ability of information understanding. In many image-text tasks, it surpasses the old single-modal scheme and shows strong transfer ability. Moreover, a single multimodal pre-trained model can be directly used to handle multimodal tasks, simplifying the integration and management of the system. Therefore, this study is the first to use the multi-modal pre-trained model CLIP in the field of intent recognition to extract the features of text and images and achieve multi-modal representation.

CLIP (Contrastive Language-Image Pre-training) model is a multimodal pre-training model developed by OpenAI based on 400 million image-text data pairs. It performs well in text and image processing tasks and achieves state of the art performance (SOTA) in many tasks. It uses a contrastive learning method for pre-training, which maps images and text to a common embedding space by maximizing the similarity between relevant image and text pairs while minimizing the similarity between irrelevant image and text pairs, which enables CLIP to understand text and images simultaneously. CLIP is pre-trained on a large-scale multi-modal data set. This large-scale data set helps the model learn more general features and can also be fine-tuned on specific tasks to adapt the model to specific fields or applications, thus having versatility and portability, and being able to adapt to different application scenarios. As shown in the figure below, CLIP mainly consists of two parts: Text Encoder and Image Encoder. Text Encoder is used to extract text features and can use the masked self-attention Transformer common in NLP; while Image Encoder is used to extract image features and can Adopts the latest proposed ViT-B/32 Transformer architecture.

ViT-B/32 Transformer architecture is used for image coding. It is an image classification model based on Transformer, where ViT represents Vision Transformer, B represents the basic version, and 32 represents that the image is divided into 32×32 image blocks. Compared with traditional convolutional neural networks (CNN), the ViT model adopts a pure Transformer structure, treating images as a series of patch sequences for processing, and has better global perception capabilities and generalization performance. In addition, the ViT model also has the advantage of being highly scalable and can improve performance by increasing the depth and width of the model.

Masked Self-Attention Transformer is a deep learning method based on Transformer architecture, which is mainly used to process sequence data in text and has strong representation ability and generalization ability. By adopting the Masked Self-Attention mechanism, it enables the model to focus on different parts in the input sequence and generate corresponding outputs based on context information.

In the intention detection task based on image and text fusion, in addition to extracting the features of different modalities, it is more important to fuse the features of different modalities. Multi-modal feature fusion is an important process for the model to integrate multiple modalities for prediction tasks. Due to the complementarity and difference between different modal data, the contribution to the results is also different. Feature fusion can provide more effective information for model prediction and improve the accuracy of prediction.

Currently, common multi-modal fusion strategies are feature-level fusion [31] , decision-level fusion [32] and hybrid fusion [33]. Decision-level fusion can use suitable models for training for different modalities, so it can better extract the internal information of a single modality and has good generalization. However, each modality uses different models for training, which cannot well capture the interaction information between different modalities and is easy to ignore the correlation between different modalities. The hybrid fusion method is flexible in design and has the advantages of both feature-level fusion and decision-level fusion. However, this method is relatively complex, difficult to implement, can easily cause over-fitting problems, and is suitable for scenarios with three modalities and above.

In order to extract deep features of different modalities and better integrate information between different modalities, this paper adopts a feature-level fusion strategy to fuse image and text features based on a cross-modal attention mechanism. Different from the simple vector splicing method, based on Multi-modal fusion with cross-modal attention mechanism refers to using the attention mechanism to dynamically adjust the attention between modalities when processing multi-modal data to achieve more effective information fusion. In multi-modal fusion, the cross-modal attention mechanism allows the model to dynamically adjust the attention to different modalities at each moment, capturing important features while excluding noise. In this way, the model can better understand the overall structure of the multimodal data, thereby improving the performance of the task.

We input the vector obtained by the fusion layer into the multi-layer perceptron. For the intention recognition in this article, it is essentially a multi-classification problem. SoftMax can be used as the last layer of the neural network to calculate the intention prediction score. SoftMax is an activation function that normalizes a numeric vector into a probability distribution vector, and the sum of each probability is 1.

where W and b represent linear layer parameters and bias terms respectively. Using Cross Entropy as the loss function, Cross Entropy is an important concept in Shannon information theory and is mainly used to measure the difference information between two probability distributions.

n is the total number of intentions, y is the one-hot representation of the sample label, and y represents the probability that the sample belongs to the i-th category.