



OPEN ACCESS

EDITED BY

Xianmin Wang,
Guangzhou University, China

REVIEWED BY

Tao Han,
Hubei Normal University, China

Fei Yan,
Changchun University of Science and
Technology, China

*CORRESPONDENCE

An Jianliang
✉ ajl13832187441@163.com

RECEIVED 22 June 2024

ACCEPTED 24 September 2024

PUBLISHED 31 October 2024

CITATION

Jianliang A (2024) A multimodal educational
robots driven via dynamic attention.
Front. Neurobot. 18:1453061.
doi: 10.3389/fnbot.2024.1453061

COPYRIGHT

© 2024 Jianliang. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC
BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

A multimodal educational robots driven via dynamic attention

An Jianliang^{1,2*}

¹College of Education, Hebei Normal University, Hebei, China, ²Industrial Fundamentals Teaching
Department, Hebei Vocational University of Industry and Technology, Shijiazhuang, Hebei, China

Introduction: With the development of artificial intelligence and robotics technology, the application of educational robots in teaching is becoming increasingly popular. However, effectively evaluating and optimizing multimodal educational robots remains a challenge.

Methods: This study introduces Res-ALBEF, a multimodal educational robot framework driven by dynamic attention. Res-ALBEF enhances the ALBEF (Align Before Fuse) method by incorporating residual connections to align visual and textual data more effectively before fusion. In addition, the model integrates a VGG19-based convolutional network for image feature extraction and utilizes a dynamic attention mechanism to dynamically focus on relevant parts of multimodal inputs. Our model was trained using a diverse dataset consisting of 50,000 multimodal educational instances, covering a variety of subjects and instructional content.

Results and discussion: The evaluation on an independent validation set of 10,000 samples demonstrated significant performance improvements: the model achieved an overall accuracy of 97.38% in educational content recognition. These results highlight the model's ability to improve alignment and fusion of multimodal information, making it a robust solution for multimodal educational robots.

KEYWORDS

ALBEF, VVG19, dynamic attention mechanism, multimodal robot, educational

1 Introduction

In recent years, the application of multimodal educational robots in the field of education has received widespread attention. Multimodal educational robots not only enhance students' learning experiences through various senses such as vision and hearing but also improve students' learning motivation and engagement through human-machine interaction (Liang and Hwang, 2023). Additionally, multimodal educational robots can integrate artificial intelligence technologies to achieve personalized teaching and intelligent tutoring, thereby better meeting the learning needs of different students (Lin et al., 2022). Therefore, researching the application and development of multimodal educational robots not only helps improve the quality and efficiency of education but also promotes innovation and advancement in educational technology (Tozadore and Romero, 2020).

Traditional approaches primarily rely on symbolic AI and knowledge representation to implement the functions of multimodal educational robots. Firstly, expert systems simulate the decision-making process of human experts by encoding their knowledge. They are highly interpretable and can provide explicit reasoning for each recognition result. For example, Cheng et al. (2022) proposed a knowledge-based expert system for human action analysis in educational robots. Additionally, a comprehensive review by Zhao (2023) showcases various applications and developments of expert systems in the field of education. Secondly, rule-based methods utilize a set of predefined rules for action recognition and interaction of educational robots. These methods exhibit high determinism and reliability, performing well in complex or dynamic educational scenarios.

Lin and Zou (2022) introduced a rule-based system for automated educational analysis, while Xu and Yan (2020) presented a rule-based framework for analyzing student learning performance. Furthermore, logistic regression, as a statistical method, learns features from training data to make classification decisions. It has important applications in action recognition for educational robots and significantly improves classification accuracy. Sun and Ma (2021) demonstrated the application of logistic regression in educational robots, while Pang (2022) further investigated action analysis in educational science, using logistic regression models to enhance recognition accuracy. These methods offer advantages such as strong interpretability and transparency in the decision-making process. However, these methods have limitations in handling complex and dynamic actions, as well as limited capabilities in processing large-scale data.

To address the limitations of interpretability and performance in handling complex and dynamic actions, data-driven and machine learning algorithms have been applied to multimodal educational robots, primarily focusing on learning features from large amounts of data and performing pattern recognition. These approaches offer advantages such as automation, efficiency, and high accuracy. Firstly, Support Vector Machines (SVM) are widely used for classification tasks and can effectively handle high-dimensional data. Ng et al. (2020) demonstrated the application of SVM in educational robots, achieving precise recognition of student behaviors through accurate classification. Additionally, Sarker (2023) proposed an SVM-based system for student emotion detection, significantly improving the accuracy of emotion recognition. Secondly, Random Forest, as an ensemble learning method, improves classification stability and accuracy by constructing multiple decision trees. Ossai (2019) showed the effectiveness of Random Forest in predicting student performance. Another study by Yağcı (2022) showcased the strong performance of Random Forest in educational data analysis. Lastly, Multilayer Perceptron (MLP), as a neural network structure, can learn complex nonlinear relationships. Mamatnabiyev et al. (2024) demonstrated the superiority of MLP in predicting student grades. Furthermore, Deyzel and Theart (2023) validated the effectiveness of MLP in educational robots, significantly enhancing the robot's understanding and responsiveness to student behaviors. However, these methods have limitations in their high dependency on large-scale data and long model training times.

To address the limitations of statistical and machine learning algorithms in handling complex and dynamic actions, deep learning-based algorithms have been applied to multimodal educational robots. These algorithms primarily rely on extracting high-level features and performing pattern recognition from large datasets. This approach offers advantages such as automation, efficiency, and high accuracy. Firstly, Convolutional Neural Networks (CNNs) have been widely used for image and video processing, enabling high-precision action recognition through hierarchical feature extraction. Yoshizawa et al. (2020) utilized CNN models to analyze students' facial expressions for estimating confusion levels. Additionally, Robinson and Nejat (2023) research showcased the application of CNNs in

multimodal inputs within a home environment, supporting the daily activities of older adults. Secondly, reinforcement learning has significantly improved the autonomy and adaptability of educational robots by learning optimal strategies through interaction with the environment. Elgajji (2015)'s research demonstrated the application of reinforcement learning in multimodal robot communication systems. Yang et al. (2021) compared multiple deep learning methods for multimodal anomaly detection and found that reinforcement learning models based on Long Short-Term Memory (LSTM) performed best in terms of efficiency and accuracy. Lastly, Transformers, as models based on attention mechanisms, achieve efficient information processing by capturing global dependencies. Ye et al. (2023) proposed a real-time object detection method based on Transformers. Ortega and Faisal (2021)'s research showcased the application of attention mechanisms in multimodal signal fusion, improving the decoding of EEG and fNIRS signals for handgrip force estimation. However, these approaches suffer from the limitations of high dependency on large-scale data and high computational complexity.

This paper proposes a method that combines ALBEF (Align before Fuse), VGG19 (Visual Geometry Group), and a dynamic attention mechanism to evaluate and optimize the educational capabilities of multimodal robots. Firstly, we use the ALBEF method to preprocess visual and textual data to enhance the understanding and alignment of visual and textual information. ALBEF improves the matching and consistency of multimodal information by aligning visual and textual data before fusion. Then, we introduce the VGG19 model as a feature extractor to extract feature representations of image data. VGG19 is a classic convolutional neural network architecture known for its excellent performance in image classification and feature extraction. Next, we employ a dynamic attention mechanism to further enhance the perception and understanding capabilities of multimodal information. The dynamic attention mechanism can automatically learn the key parts of the input data and focus on important information, thereby improving the model's performance and effectiveness. By integrating ALBEF, VGG19, and the dynamic attention mechanism, we construct a multimodal robot education model and perform training and optimization.

To address the challenges of high computational complexity and long model training time, this paper proposes Res-ALBEF: a dynamic attention-driven multimodal educational robot that aims to address the alignment and processing of multimodal data in educational robots. Firstly, the ALBEF method preprocesses visual and textual data to enhance their understanding and alignment, thus improving the consistency of multimodal information. Secondly, the VGG19 model serves as a feature extractor for image data. Lastly, the dynamic attention mechanism automatically identifies and focuses on the important parts of the input data, thereby improving the overall effectiveness of the model. By integrating ALBEF, VGG19, and the dynamic attention mechanism, Res-ALBEF significantly enhances the educational capabilities of multimodal robots.

The three contributions of this paper are as follows:

- The proposed method introduces a novel multimodal network architecture called Res-ALBEE, which presents a new approach for the alignment and processing of multimodal data in educational robots.
- This method is efficient and versatile, suitable for various educational scenarios, ensuring robust performance across different tasks, and environments.
- Experimental results demonstrate a significant enhancement in the educational capabilities of multimodal robots, showcasing improved alignment, feature extraction, and attention mechanisms.

2 Related work

2.1 Multimodal perception and understanding

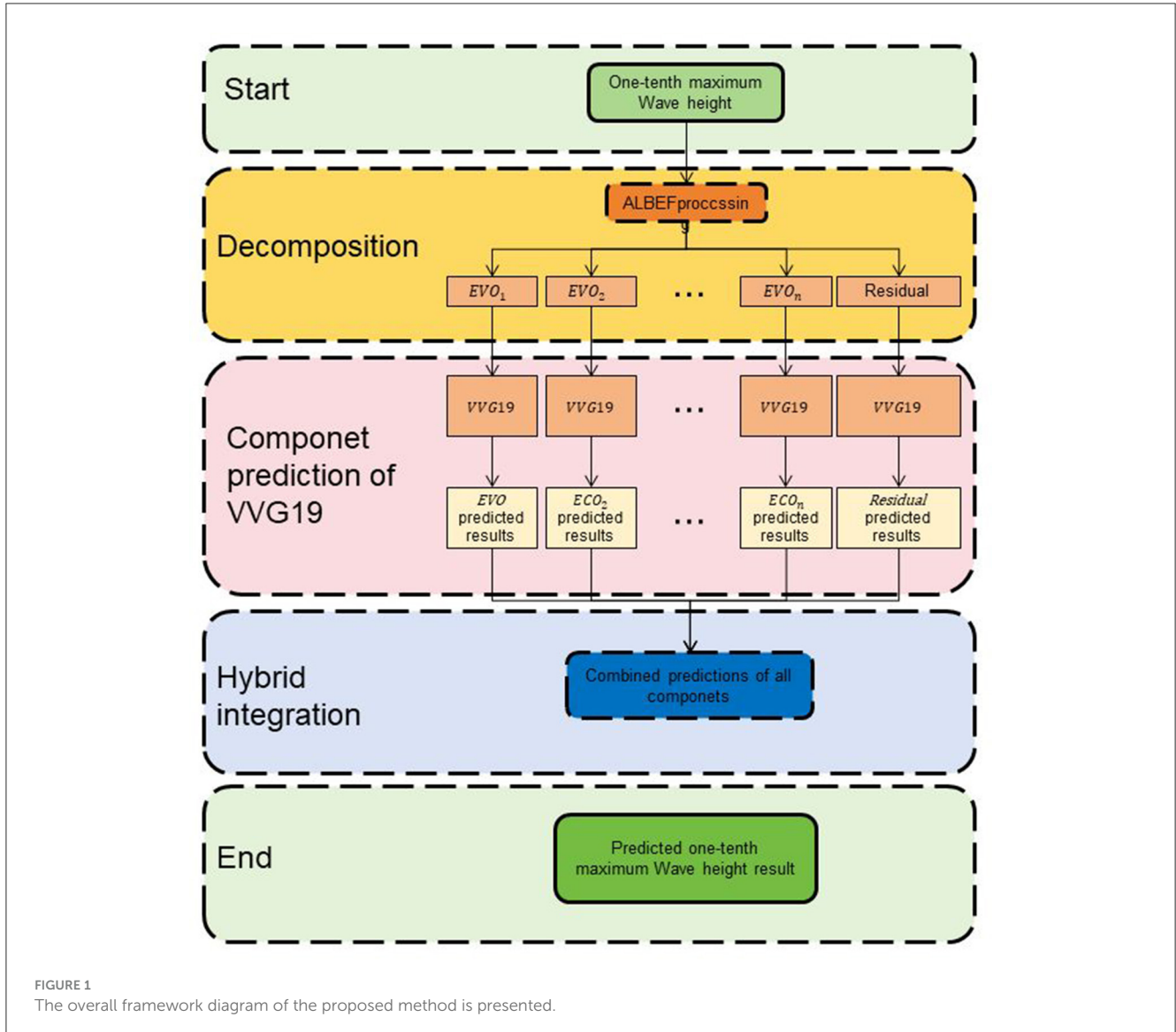
Multimodal perception and understanding is a key direction in the field of multimodal robot education. By integrating various sensory modalities such as vision, hearing, and touch, multimodal robots can more comprehensively perceive the learning environment and the state of the students, thus providing a richer and more personalized educational experience (Braud et al., 2020). Researchers face many challenges in this direction, including the alignment of multimodal data, feature extraction and representation learning, and how to use multimodal information to understand and generate educational content (Hong et al., 2024). Firstly, the alignment of multimodal data is a significant issue. Data from different sensory modalities have different formats and characteristics, such as images, speech, and text. Researchers need to explore effective methods to align these data so that information from different modalities can match and integrate. For example, alignment networks or adversarial generative networks can be used to achieve alignment and conversion between modalities (Yan et al., 2021). Secondly, feature extraction and representation learning are critical tasks for multimodal perception and understanding. Researchers need to design deep learning models or neural network architectures to extract meaningful feature representations from multimodal inputs. These features should capture the correlations and semantic information between different modalities, providing useful inputs for subsequent educational tasks. For instance, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can be used for feature extraction and sequence modeling. Furthermore, an important area of research is the utilization of multimodal information for comprehending and creating educational content. Researchers can investigate methods for combining visual, auditory, and textual information to enhance the understanding and context of educational materials. For example, attention mechanisms can be used to automatically learn the key parts of input data and focus on important information. Additionally, research can be conducted on how to generate multimodal educational responses, such as image descriptions and speech answers, to provide effective interaction capabilities with students (Lin et al., 2024).

2.2 Personalized and adaptive learning

Personalized and adaptive learning is another critical direction in the field of multimodal robot education. Traditional educational models often adopt a one-size-fits-all approach, which fails to meet the individual differences and learning needs of students. In contrast, multimodal robot education systems, equipped with perception and interaction capabilities, can provide personalized and adaptive educational services based on individual student characteristics, learning progress, and emotional states (Yan et al., 2022). Firstly, personalized education requires the use of multimodal robots to perceive and analyze students' individual characteristics. For instance, through visual perception, robots can recognize students' facial expressions and body language to understand their emotional states and levels of attention. Through auditory perception, robots can analyze students' voice features and speech rates, assessing pronunciation accuracy and fluency. Subsequently, multimodal robots need to design and optimize educational strategies and interaction methods based on students' individual characteristics and learning progress. For students with slower learning progress, robots can adopt more detailed and gentle explanations, providing additional support materials and practice opportunities. For students with faster learning progress, robots can offer more advanced and challenging learning content to stimulate their interest and motivation (Wang et al., 2021). Additionally, adaptive learning is a crucial component of personalized education. Multimodal robots can adjust teaching strategies and content in real time based on students' learning behaviors and feedback to provide the most suitable learning paths and educational resources. For example, robots can offer targeted feedback and tutoring based on students' incorrect answers and error patterns, helping them overcome difficulties and improve learning outcomes (Hong et al., 2024).

2.3 Human-robot collaboration and interactive learning

Human-robot collaboration and interactive learning is another important research direction in the field of multimodal robot education. Multimodal robots, acting as educational assistants and partners, need to engage in effective human-robot interactions and collaborations with students to facilitate learning and knowledge transfer. Research in this direction aims to design the expressions and interaction methods of robots so that they can communicate with students in a natural, intelligent, and effective manner (Wang et al., 2019). Firstly, researchers can explore how to design the language and non-verbal expression capabilities of robots. Robots need to have good language understanding and generation abilities to converse and interact with students. For example, robots can understand students' questions and needs, and respond appropriately, providing targeted explanations and guidance. Additionally, robots can interact with students through facial expressions, gestures, and postures, enhancing the richness of communication and emotional connection (Bera et al., 2020). Secondly, researchers need to focus on the cooperation and

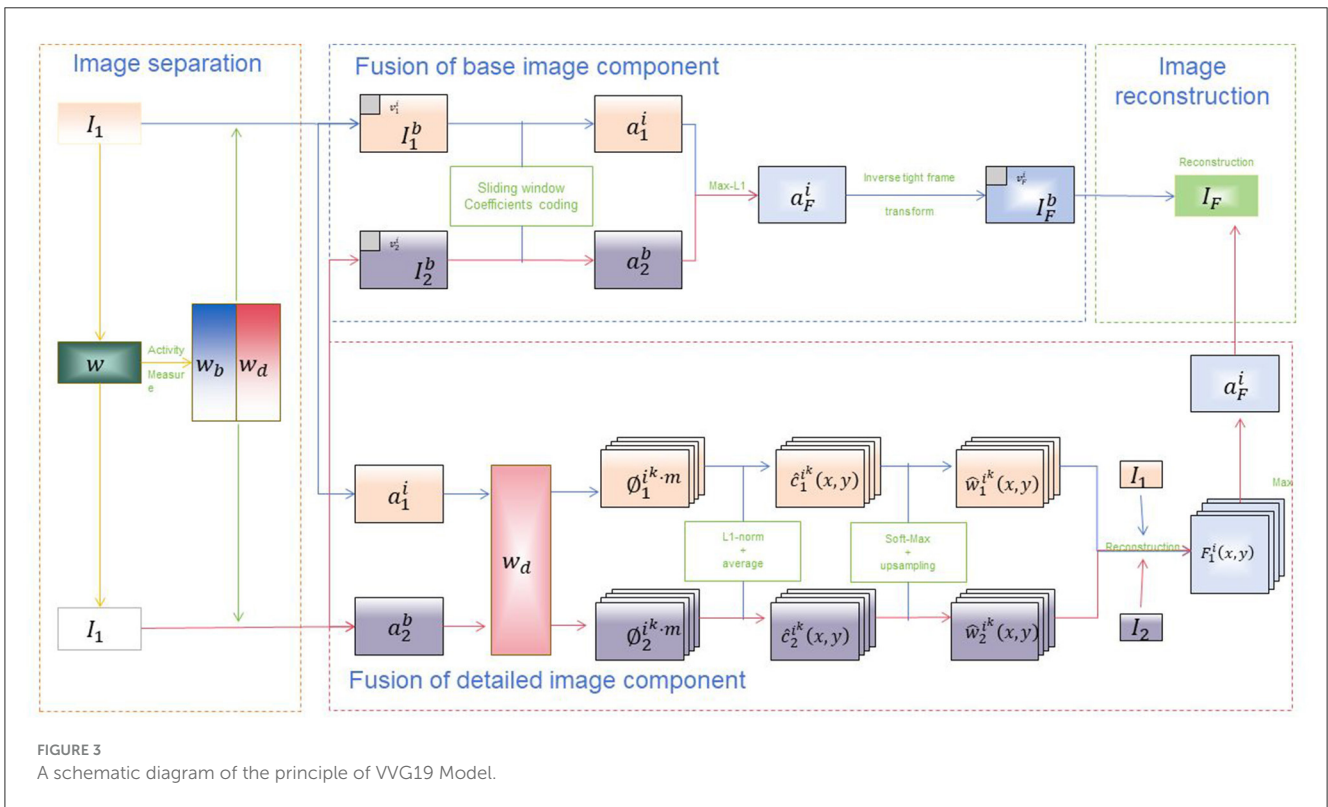
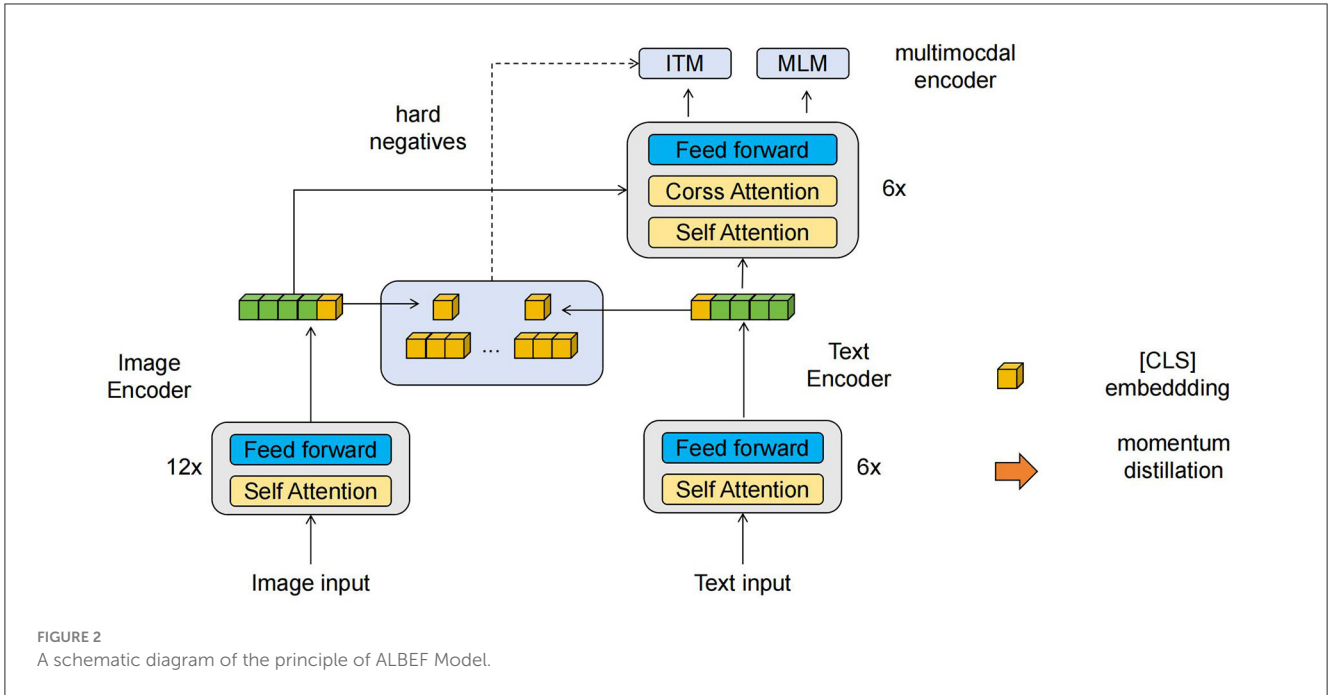


collaborative learning between robots and students. Multimodal robots can act as learning partners, working with students to solve problems, discuss learning content, and complete tasks and projects together (Ai et al., 2023). Through collaboration with robots, students can gain practical experience and develop hands-on skills, teamwork, and problem-solving abilities. Researchers can study how to design the cooperation strategies and interaction modes of robots to achieve efficient human-robot collaboration and interactive learning. Moreover, human-robot collaboration involves the robot's ability to adapt to the student's learning pace and style, providing personalized assistance and feedback. Robots can monitor students' progress and adjust their support accordingly, ensuring that students remain engaged and motivated. For instance, when a student struggles with a concept, the robot can offer additional explanations or alternative approaches, fostering a supportive learning environment (Lazaro et al., 2021).

3 Methodology

3.1 Overview of our network

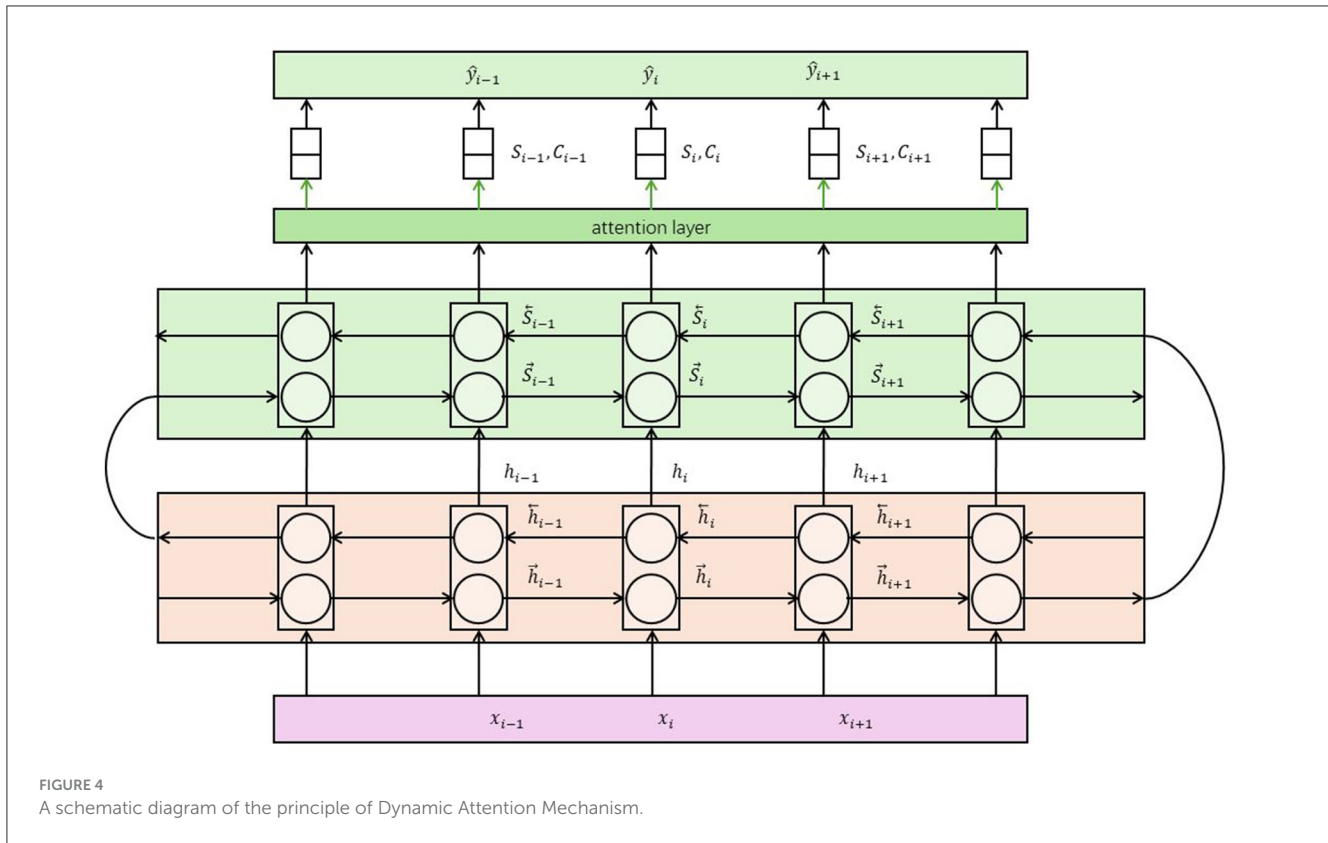
This study aims to evaluate and optimize the educational capabilities of multimodal robots using Res-ALBEF: A Multimodal Educational Robot driven by Dynamic Attention. Res-ALBEF combines ALBEF (Align Before Fuse) and VGG19 with a dynamic attention mechanism. ALBEF aligns text and visual representations to capture semantic relationships, while VGG19 extracts visual features from images or videos. The dynamic attention mechanism enables the model to focus on relevant parts of the input data based on task relevance. Figure 1 shows the overall framework diagram of the proposed method.



3.2 Model architecture

Design a multimodal architecture that integrates ALBEF, VGG19, and a dynamic attention mechanism. This model should be capable of aligning text and visual representations, using VGG19 to extract visual features, and dynamically focusing on relevant information during educational tasks. In order to evaluate

the performance of the multimodal educational robot model, we defined appropriate evaluation indicators. These indicators include: Accuracy, which evaluates the accuracy of the overall prediction of the model; Precision, which evaluates the proportion of correctly predicted positive samples among all predicted positive samples; Recall, which evaluates the proportion of correctly predicted positive samples among all actually positive samples;



F1 Score, the harmonic mean of precision and recall, which comprehensively measures model performance. In addition, we also use the educational content understanding score to evaluate the accuracy and depth of the robot’s understanding of educational content, and the response generation score to evaluate the relevance and effectiveness of the robot’s generated responses in educational scenarios.

The main reason for combining ALBEF, VGG19, and dynamic attention mechanisms is that these three components bring complementary strengths to the processing of multimodal data. By working together, they overcome the current limitations in multimodal systems, particularly in data alignment, feature extraction, and information focus. This synergy enhances the overall performance of the system. ALBEF (Align Before Fuse) plays a crucial role by ensuring that different modalities, visual and textual information are aligned before being fused. In multimodal tasks, visual, and textual data often have distinct representations, and fusing them prematurely can lead to mismatched or lost information. ALBEF provides a mechanism to align these modalities in a high-dimensional feature space, enabling better understanding and interaction between them. This alignment is particularly critical for educational robots, where student behaviors and expressions (visual data) must be accurately linked with course content (textual data) to enable the robot to respond appropriately.

VGG19, a powerful image feature extractor with a deep convolutional structure, extracts fine-grained visual features, especially in complex visual inputs. It captures multi-level image

information such as edges, shapes, and objects, providing a strong foundation for handling visual data. In scenarios where educational robots need to analyze visual scenes or student facial expressions, VGG19 ensures the extraction of highly accurate visual features, which, when combined with ALBEF’s alignment process, leads to more efficient and precise multimodal fusion. Finally, the dynamic attention mechanism dynamically adjusts the model’s attention based on the characteristics of the input data, focusing computational resources on the most relevant information. This is essential for multimodal data processing, as only portions of the visual or textual input may be highly relevant to the current task. The dynamic attention mechanism filters out irrelevant noise, allowing the model to focus on key aspects of the data, thereby improving processing efficiency and accuracy. Together, these three components form a cohesive system: ALBEF ensures proper alignment of the modalities, VGG19 provides rich visual features, and the dynamic attention mechanism focuses on the most critical information, creating a robust, flexible, and efficient framework for multimodal data processing, particularly in educational robotics.

3.3 ALBEF

The ALBEF (Align before Fuse) (Li et al., 2021) model’s fundamental principle is to align textual and visual representations before fusing them in multimodal tasks (Zeng et al., 2021). This

TABLE 1 Performance comparison on MS COCO and Ref COCO datasets.

| Model | Datasets | | | | | | | |
|--------------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | MS COCO dataset | | | | Ref coco dataset | | | |
| | Accuracy | Recall | F1 score | AUC | Accuracy | Recall | F1 score | AUC |
| Ionescu (2022) | 88.33 ± 0.03 | 89.65 ± 0.02 | 91.14 ± 0.01 | 92.61 ± 0.02 | 87.6 ± 0.03 | 86.47 ± 0.02 | 89.13 ± 0.01 | 92.39 ± 0.02 |
| González Ybarra (2022) | 90.16 ± 0.03 | 91.9 ± 0.02 | 88.29 ± 0.01 | 92.26 ± 0.02 | 86.72 ± 0.03 | 93.66 ± 0.02 | 90.32 ± 0.01 | 85.48 ± 0.02 |
| Ptak et al. (2024) | 87.49 ± 0.03 | 87.33 ± 0.02 | 84.47 ± 0.01 | 93.06 ± 0.02 | 92.83 ± 0.03 | 86.48 ± 0.02 | 90.39 ± 0.01 | 86.9 ± 0.02 |
| Khilji et al. (2021) | 96.14 ± 0.03 | 91.78 ± 0.02 | 90.04 ± 0.01 | 91.02 ± 0.02 | 95.62 ± 0.03 | 84.42 ± 0.02 | 90.19 ± 0.01 | 86.95 ± 0.02 |
| Li (2023) | 93.66 ± 0.03 | 93.04 ± 0.02 | 90.53 ± 0.01 | 90.21 ± 0.02 | 91.19 ± 0.03 | 88.93 ± 0.02 | 88.45 ± 0.01 | 90 ± 0.02 |
| Minoli and Occhiogrosso (2020) | 90.69 ± 0.03 | 92.99 ± 0.02 | 90.15 ± 0.01 | 87.23 ± 0.02 | 95.57 ± 0.03 | 88.54 ± 0.02 | 90.35 ± 0.01 | 93.22 ± 0.02 |
| Ours | 97.38 ± 0.03 | 94.35 ± 0.02 | 92.84 ± 0.02 | 95.58 ± 0.02 | 97.63 ± 0.03 | 95.15 ± 0.02 | 93.72 ± 0.02 | 96.69 ± 0.02 |

Bold font indicates the best value.

method aims to capture the semantic relationships between text and visual data to enhance the understanding of multimodal information. Figure 2 is a schematic diagram of the principle of ALBEF Model.

In multimodal tasks, there often exists an information mismatch between text and visual data. For example, when describing an image, the text can provide a semantic description of the image’s content, while visual information offers more intuitive visual features. ALBEF’s goal is to find a way to align these two types of representations to better fuse them and improve the understanding of multimodal data. The key function of ALBEF is to achieve the alignment of textual and visual representations. The specific steps can include the following aspects: Textual and Visual Feature Extraction: First, extract feature representations from text and visual data. For text data, natural language processing techniques such as word embeddings or text encoders can be used to convert the text into vector representations. For visual data, methods such as convolutional neural networks (CNNs) can be used to extract visual features from images or videos. Cross-modal Alignment: In ALBEF, a cross-modal alignment method is used to align textual and visual representations. This can be achieved through cross-modal retrieval or cross-modal attention mechanisms.

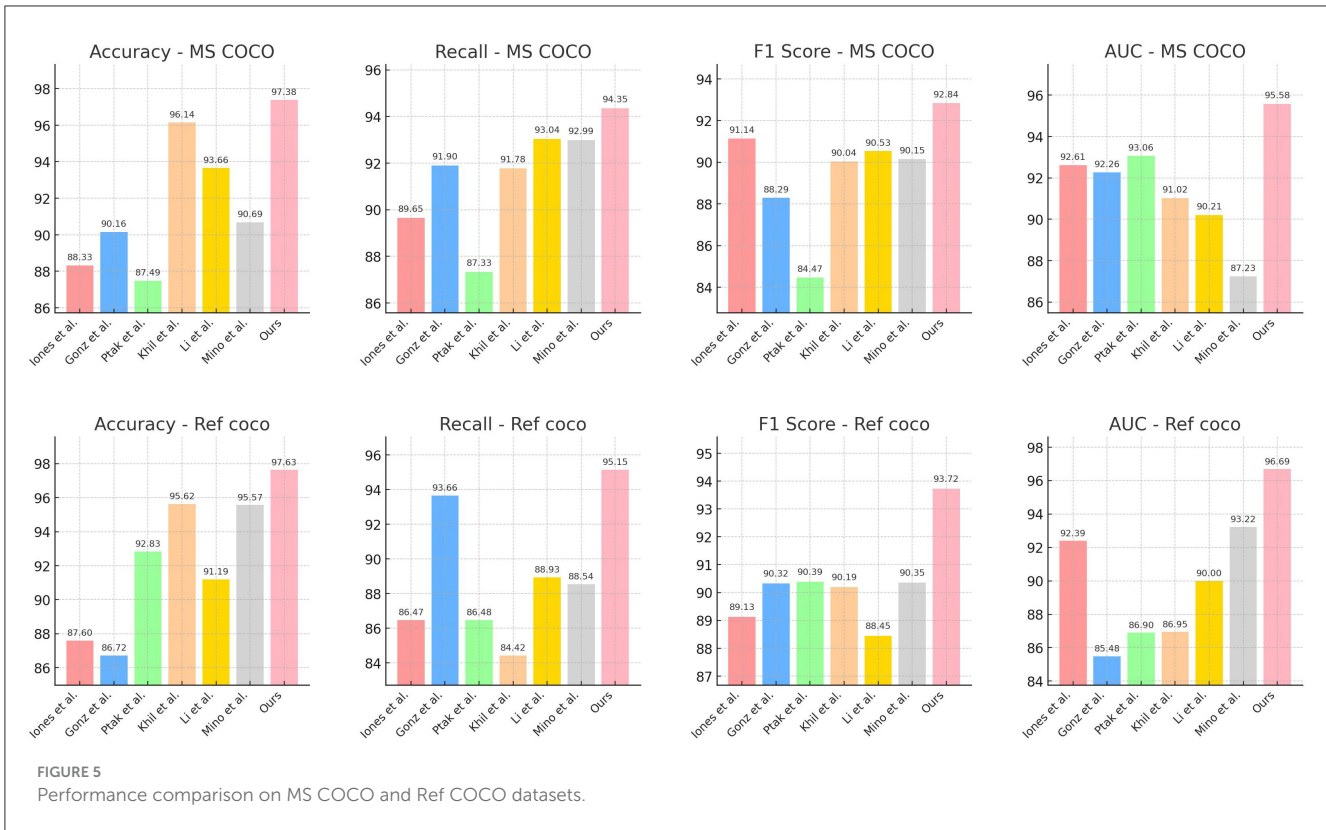
In terms of cross-modal retrieval, the most relevant text-visual pairs are identified by calculating the similarity or distance between text and visual representations. We use the following metrics to measure the similarity between text and visual data: Cosine Similarity, which measures the cosine angle between two vectors, where higher similarity means the two vectors are closer; Euclidean Distance, which measures the straight-line distance between two vectors, where smaller distance means the two vectors are closer. Cross-modal Attention: By computing attention weights between textual and visual representations, it dynamically focuses on the relevant textual and visual information. This can be achieved by introducing attention mechanisms such as bilinear attention or adaptive attention. Aligned Feature Fusion: After aligning textual and visual representations, the aligned features are fused. Fusion can be done through simple concatenation, weighted summation, or using neural network models to learn fusion weights. The key idea of ALBEF is to address the information mismatch problem in multimodal tasks by aligning textual and visual representations before fusion. Through alignment, ALBEF can better capture the semantic relationships between text and visual data, thereby enhancing the understanding of multimodal information. In the context of multimodal robot education tasks, ALBEF enhances the robot’s ability to understand and respond to educational content, improving its performance in the education domain.

The formula of ALBEF (Align before Fuse) is as follows (Equation 1):

$$\text{Align}(T, V) = \text{softmax}\left(\frac{T \cdot V^T}{\sqrt{d}}\right) \cdot V \quad (1)$$

Where the variables have the following meanings:

T : text representation, dimension is $n \times d$, where n is the number of text samples and d is the text feature dimension. V : visual representation, dimension is $m \times d$, where m is the number of visual samples and d is the visual feature dimension.



Align(T, V): aligned visual representation, dimension is $n \times d$, the same as the text representation. softmax(\cdot): softmax function, used to calculate the attention weight so that it satisfies the properties of probability distribution. $\frac{T \cdot V^T}{\sqrt{d}}$: The similarity matrix between text and vision, obtained by calculating the dot product of the text representation and the visual representation, and normalized by \sqrt{d} . \cdot : Matrix multiplication operation. This formula represents the alignment operation in the ALBEF model. The attention weights are obtained by calculating the similarity matrix between text and vision and normalizing it through the softmax function. Then, the attention weights are multiplied with the visual representation to obtain the aligned visual representation. In this way, the text and visual representations are semantically aligned.

3.4 VVG19

VGG19 (Effati and Nejat, 2023) proposed by the Visual Geometry Group, is a deep convolutional neural network model designed to extract visual features from images through operations such as convolution and pooling (Rajangam et al., 2021). It is one of the models in the VGG series, consisting of 19 layers of convolutional and fully connected layers, featuring a deep network structure. Figure 3 is a schematic diagram of the principle of VVG Model.

Basic Principles of VGG19: Preprocessing of Input Images: VGG19 first preprocesses the input images, which includes

normalizing pixel values, subtracting the mean, or using other preprocessing methods. Convolutional and Pooling Layers: VGG19 consists of multiple convolutional layers and pooling layers. These stacked layers extract features at different levels of abstraction from the images. VGG19 uses small-sized (3×3) convolutional kernels and (2×2) pooling kernels to increase the network's depth and non-linear capability. Learning Feature Representations: By adding activation functions (such as ReLU) after convolutional layers, non-linearity is introduced, enabling the network to learn feature representations from images. Through the stacking of multiple convolutional and pooling layers, VGG19 gradually learns higher-level feature representations, such as edges, textures, and parts of objects. Fully Connected Layers: After the convolutional and pooling layers, VGG19 includes several fully connected layers to map the learned features to the predicted categories. The fully connected layers typically include multiple hidden layers and an output layer, where the number of nodes in the output layer corresponds to the number of categories in the task. VGG19 is commonly used as a visual feature extractor in multimodal approaches. In this study, we leverage a pre-trained VGG19 model to extract rich semantic visual features from images or videos. These features are solely used for visual feature extraction and integrated with textual or other modal features to enhance the model's understanding of multimodal data.

The VGG19 (Visual Geometry Group 19) model does not have a single formula to describe it because it is a deep neural network consisting of multiple convolutional layers, pooling layers, and fully

TABLE 2 Performance comparison on MS COCO and Ref COCO datasets.

| | CC12M Dataset | | | | VG-Cap Dataset | | | |
|---------------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| | Parameters (M) | Flops (G) | Inference time (ms) | Training time (s) | Parameters (M) | Flops (G) | Inference time (ms) | Training time (s) |
| Ionescu (2022) | 282.98 ± 0.03 | 275.62 ± 0.03 | 243.77 ± 0.03 | 214.42 ± 0.03 | 250.34 ± 0.03 | 318.30 ± 0.03 | 374.37 ± 0.03 | 267.89 ± 0.03 |
| González Ybarra (2022) | 252.43 ± 0.03 | 260.36 ± 0.03 | 237.45 ± 0.03 | 232.15 ± 0.03 | 212.41 ± 0.03 | 233.57 ± 0.03 | 364.68 ± 0.03 | 231.49 ± 0.03 |
| Prak et al. (2024) | 265.03 ± 0.03 | 271.39 ± 0.03 | 344.88 ± 0.03 | 208.29 ± 0.03 | 295.45 ± 0.03 | 313.00 ± 0.03 | 201.70 ± 0.03 | 356.11 ± 0.03 |
| Khilji et al. (2021) | 310.34 ± 0.03 | 200.64 ± 0.03 | 365.81 ± 0.03 | 266.66 ± 0.03 | 242.72 ± 0.03 | 340.73 ± 0.03 | 290.87 ± 0.03 | 264.82 ± 0.03 |
| Li (2023) | 385.22 ± 0.03 | 278.57 ± 0.03 | 343.39 ± 0.03 | 213.49 ± 0.03 | 368.48 ± 0.03 | 273.54 ± 0.03 | 208.78 ± 0.03 | 274.82 ± 0.03 |
| (Mimoli and Occhiogrosso, 2020) | 210.06 ± 0.03 | 370.23 ± 0.03 | 207.12 ± 0.03 | 310.99 ± 0.03 | 258.30 ± 0.03 | 348.60 ± 0.03 | 289.38 ± 0.03 | 351.24 ± 0.03 |
| Ours | 231.40 ± 0.03 | 149.83 ± 0.03 | 116.12 ± 0.03 | 180.31 ± 0.03 | 222.27 ± 0.03 | 126.85 ± 0.03 | 148.55 ± 0.03 | 226.84 ± 0.03 |

Bold font indicates the best value.

connected layers. However, the overall structure of the VGG19 model can be represented as follows:

$$\begin{aligned}
 & \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Pool} \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Pool} \\
 & \quad \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Pool} \\
 & \quad \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Pool} \\
 & \quad \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Pool} \\
 & \quad \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Conv} \rightarrow \text{Pool} \\
 & \quad \rightarrow \text{FC} \rightarrow \text{FC} \rightarrow \text{FC}
 \end{aligned} \tag{2}$$

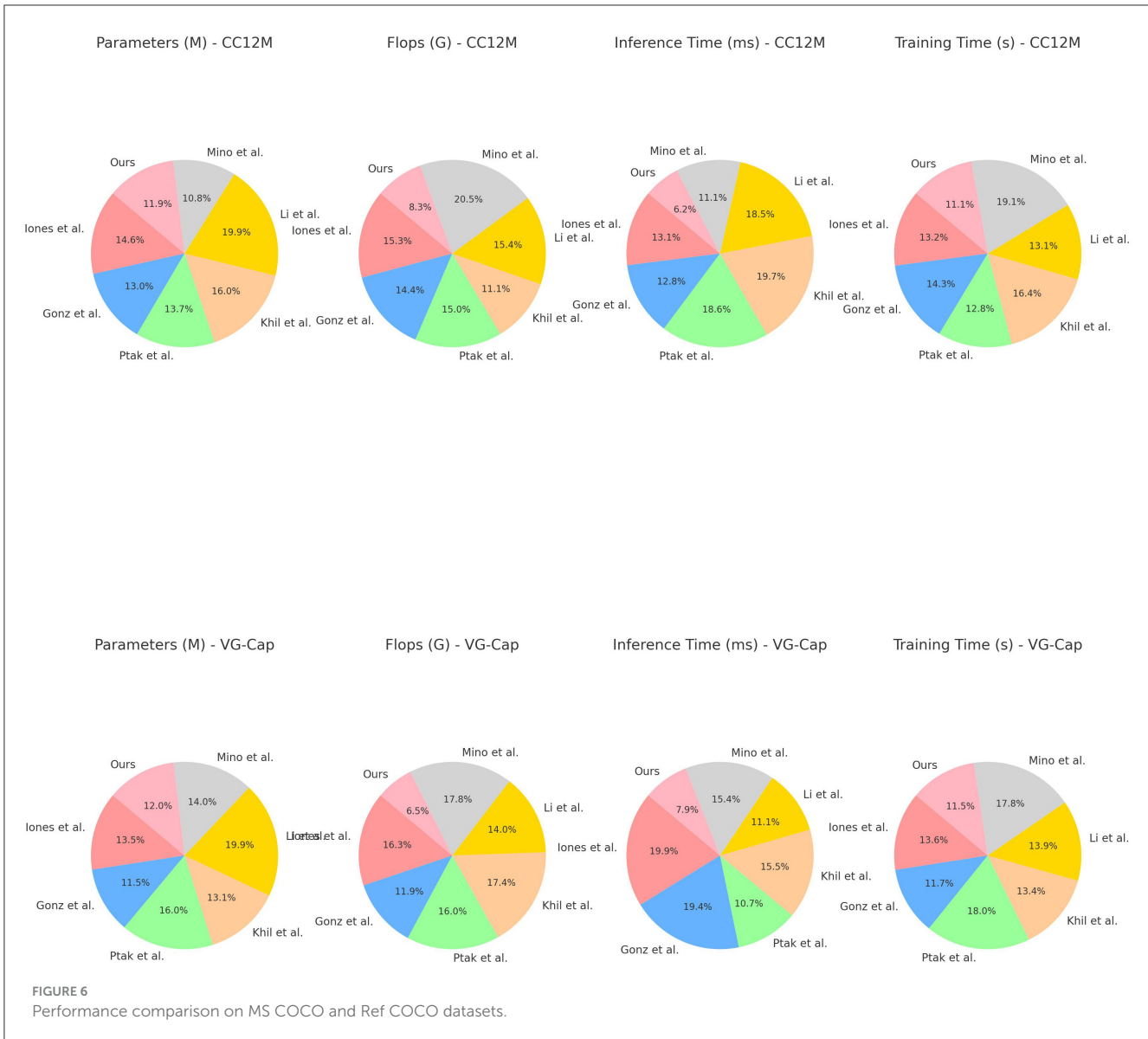
The meanings of the variables are as follows:

Conv: Convolutional layer, which uses convolution operations to extract image features. Pool: Pooling layer, which uses pooling operations to reduce the size of feature maps. FC: Fully connected layer, which maps features to predicted categories.

The VGG19 model consists of multiple convolutional layers and pooling layers stacked alternately, and finally uses several fully connected layers as output layers. Each convolutional layer uses convolution operations to extract image features and introduces nonlinearity through activation functions. The pooling layer is used to reduce the size of the feature map to reduce computational complexity and increase the translation invariance of the input. The final fully connected layer maps the learned features to the predicted category. This image provides a visual representation of the VGG19 architecture applied to a multimodal fusion process. The process is divided into several key stages: 1. **Image separation**: - The input image I_1 is divided into different components. - An activity measure w is applied to the image to produce w_b and w_d . 2. **Fusion of base image component**: - The base components of the images (I_1^b and I_2^b) are processed using sliding window coefficients coding. - The coded components a_1^i and a_2^i are then combined using a Max-L1 strategy to produce a_F^i . - An inverse tight frame transform is applied to a_F^i to reconstruct the base fused image component I_F^b , resulting in the intermediate fused image I_F . 3. **Fusion of detailed image component**: - The detailed components a_1^i and a_2^i are processed. - These components undergo L1-norm averaging and soft-max upsampling to produce weighted components $c_{1k}^i(x, y)$ and $c_{2k}^i(x, y)$. - The resulting components are then reconstructed into $F_1^i(x, y)$, which contributes to the final fused image I_F . 4. **Image reconstruction**: - The fused components are combined and reconstructed to form the final output image I_F . The diagram illustrates the complex process of image fusion using the VGG19 architecture, highlighting the detailed steps involved in separating, processing, and combining image components to achieve a final fused output. This architecture is essential in applications requiring high-quality image fusion, such as in multimodal educational robots driven by dynamic attention mechanisms.

3.5 Dynamic attention mechanism

The Dynamic Attention Mechanism (Van Amsterdam et al., 2022) is a commonly used attention mechanism in multimodal models, designed to weigh and focus on different parts of the input at various time steps or spatial locations (Ding et al., 2021).



Its basic principle is to dynamically compute attention weights based on the contextual information of the input, enabling the model to adaptively focus on important parts of the input. Figure 4 is a schematic diagram of the principle of Dynamic Attention Mechanism.

Steps to Calculate Attention Weights in Dynamic Attention Mechanism: Input Feature Representation: First, extract feature representations for each modality of the input data. This can be achieved using pretrained neural network models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), or Transformers. Context Representation: Next, compute the context representation based on the current state or contextual information of the model. This could be the hidden state from a previous time step or feature representations from other modalities. Attention Weight Calculation: Calculate attention weights using the context representation and input feature representation. This is typically done by computing the similarity or correlation between them. Common methods include dot-product attention, additive attention, and multi-head attention.

Weighted Aggregation: Multiply the input feature representation by the attention weights and perform a weighted sum to obtain the aggregated feature representation. This allows the model to focus on parts of the input that are relevant to the current task. Role in Multimodal Tasks The dynamic attention mechanism provides a flexible way for the model to adaptively focus on different parts of the input based on the contextual information. It can be used in various tasks such as image captioning, visual question answering, and multimodal machine translation. By using the dynamic attention mechanism, the model can selectively extract useful information from the input while ignoring noise or irrelevant parts. This helps improve the model's performance and robustness, enabling it to better understand and model the relationships in multimodal data. Additionally, the dynamic attention mechanism enhances the interpretability of the model by allowing visualization of the attention weights, which can explain the model's decisions or generated results. Translation to Multimodal Robot Education In the context of multimodal robot education, the dynamic attention mechanism allows the robot to focus on relevant parts of

TABLE 3 Ablation study on VVG19 module for performance metrics.

| | MS COCO Dataset | | | | Ref COCO Dataset | | | |
|-----------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | Accuracy | Recall | F1 score | AUC | Accuracy | Recall | F1 score | AUC |
| ResNet-50 | 88.6 ± 0.03 | 84.24 ± 0.03 | 84.83 ± 0.03 | 92.33 ± 0.03 | 95.57 ± 0.03 | 90.35 ± 0.03 | 84.42 ± 0.03 | 90.19 ± 0.03 |
| ResNet-18 | 89.13 ± 0.03 | 90.06 ± 0.03 | 87.04 ± 0.03 | 85.13 ± 0.03 | 91.99 ± 0.03 | 84.93 ± 0.03 | 88.52 ± 0.03 | 92.53 ± 0.03 |
| CNN | 89.48 ± 0.03 | 87.07 ± 0.03 | 89.52 ± 0.03 | 92.98 ± 0.03 | 89.54 ± 0.03 | 83.94 ± 0.03 | 84.02 ± 0.03 | 89.71 ± 0.03 |
| Ours | 96.95 ± 0.03 | 94.65 ± 0.03 | 92.71 ± 0.03 | 94.09 ± 0.03 | 97.23 ± 0.03 | 94.35 ± 0.03 | 92.53 ± 0.03 | 92.16 ± 0.03 |

Bold font indicates the best value.

multimodal educational content based on the current educational task. This improves the robot’s ability to understand and respond to educational content, making the interactions more effective and personalized. By visualizing attention weights, educators can also gain insights into how the robot processes information and makes decisions, further enhancing the educational experience.

$$\begin{aligned}
 \text{Context} &= \text{ComputeContext}(\text{PreviousState}) \\
 \text{AttentionWeights} &= \text{ComputeAttention}(\text{InputFeatures}, \text{Context}) \\
 \text{WeightedFeatures} &= \text{AttentionWeights} \odot \text{InputFeatures} \\
 \text{AggregatedFeatures} &= \text{Aggregate}(\text{WeightedFeatures})
 \end{aligned} \tag{3}$$

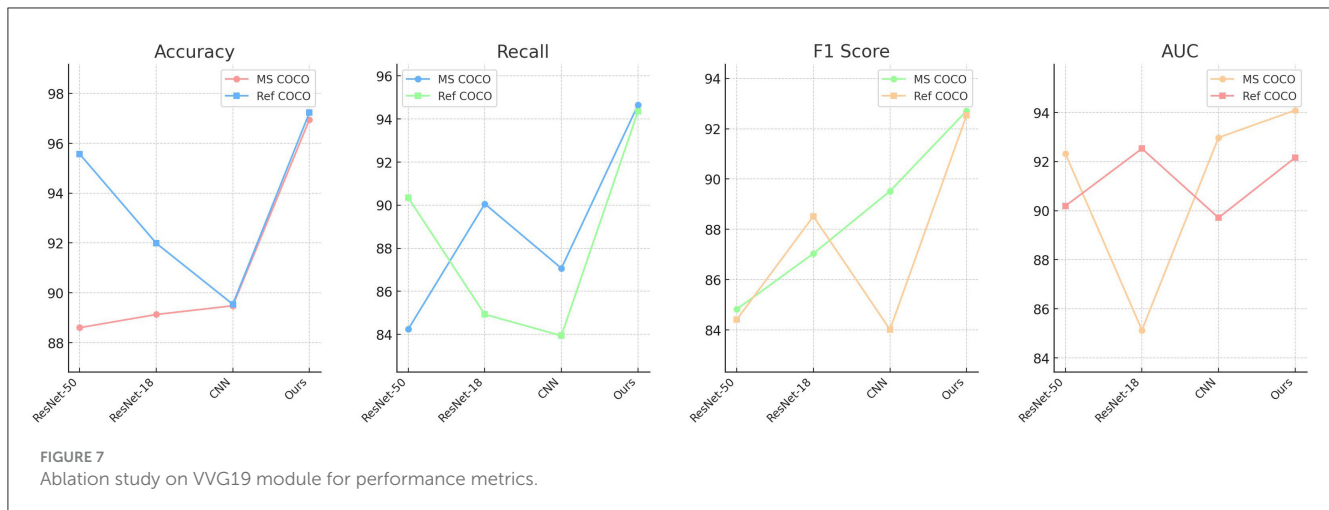
Wherein, the meanings of the variables are as follows:

Context: context representation, representation calculated based on the current state of the model or context information. PreviousState: representation of the hidden state or other modal data of the model at the previous time step. InputFeatures: input feature representation, feature representation from different modal data. AttentionWeights: Attention weights, used to weight the importance of input feature representations. \odot : Represents element-wise multiplication. WeightedFeatures: Weighted feature representation, the weighted result obtained by multiplying the input feature representation and the attention weights. AggregatedFeatures: Aggregated feature representation, the final feature representation obtained by summing or other aggregation operations on the weighted feature representations.

4 Experiment

4.1 Datasets

This paper uses four datasets: ms coco dataset, Ref coco dataset, CC12M datasets, and VG-Cap dataset. MS COCO Dataset (Tong and Wu, 2023) (Microsoft Common Objects in Context): MS COCO is a widely used dataset for image recognition and captioning tasks. It contains over 200,000 labeled images, covering a wide range of object categories. The dataset also includes human-annotated captions for each image, making it suitable for tasks related to image understanding and natural language processing. Ref COCO Dataset (Jing et al., 2021) (ReferIt Game Dataset): The Ref COCO dataset is an extension of the MS COCO dataset, specifically designed for referring expression comprehension. It contains additional annotations where human subjects provide expressions referring to specific objects in the images. This dataset is useful for tasks that involve understanding natural language references in the context of visual scenes. CC12M Datasets (Fan et al., 2024) (Conceptual Captions 12M Dataset): CC12M is a large-scale dataset consisting of 12 million image-caption pairs. The dataset emphasizes diverse and novel concepts, covering a wide range of visual and linguistic variations. It is useful for training and evaluating models in various tasks such as image captioning, image-text matching, and multimodal learning. VG-Cap Dataset (Ye and Kovashka, 2021) (Visual Genome Caption Dataset): The VG-Cap dataset contains image-caption pairs collected from the Visual Genome project. Visual Genome is a dataset that provides detailed scene understanding annotations, including



object and attribute labels, relationships between objects, and region descriptions. The VG-Cap subset focuses specifically on image captioning, making it suitable for tasks that require detailed image descriptions.

4.2 Experimental details

Experimental Design Objective: The objective of this experiment is to compare different models in terms of both performance metrics (Accuracy, AUC, Recall, F1 Score) and technical metrics (Training Time, Inference Time, Parameters, and FLOPs). **Experimental Steps:** **Dataset Selection:** Select a suitable dataset for your study, such as MS COCO, RefCOCO, CC12M, or VG-Cap. **Model Selection:** Choose several models with different architectures or characteristics for comparison. These can be classical models or the latest models applicable to the chosen dataset. Ensure that the models have varying numbers of parameters and computational complexities for a comprehensive comparison. **Experimental Setup:** **Data Splitting and Preprocessing:** Divide the dataset into training and testing sets and preprocess the data accordingly. **Model Training:** Train each model using the training set and record the training time. **Performance Metrics Comparison:** Compare the models' performance metrics: Accuracy, AUC, Recall, and F1 Score. Use charts or tables to clearly present the comparison results. **Ablation Study:** Conduct an ablation study on a high-performing model to verify the contribution of its various components. Gradually remove or modify certain components of the model, such as the attention mechanism or feature extractors, and compare the performance metrics of the modified models. This helps to evaluate the impact of each component on the model's performance. **Results Analysis:** Analyze the experimental results to compare the differences in technical and performance metrics across different models. Interpret the differences between the models and analyze their strengths and weaknesses. Evaluate the ablation study results to understand the contribution of each component to the model's performance. This structured approach allows for a comprehensive comparison

of different models, providing insights into their efficiency and effectiveness in handling the chosen dataset.

4.3 Experimental results and analysis

Table 1 and Figure 5 presents a comparison of performance metrics across different models on the MS COCO and Ref COCO datasets, including Accuracy, Recall, F1 Score, and AUC (Area Under the Curve). Our model performs best across all these metrics on both datasets, with Accuracy (97.38% and 97.63%), Recall (94.35% and 95.15%), F1 Score (92.84% and 93.72%), and AUC (95.58% and 96.69%) surpassing those of other models. This indicates that our model excels in multimodal tasks, particularly in integrating visual and textual information. By leveraging the ALBEF (Align before Fuse) framework and VVG19 (Visual Geometry Group) model, our approach first aligns visual and textual information before fusing them, thereby enhancing accuracy and consistency in information processing and understanding. Additionally, the introduction of a dynamic attention mechanism allows the model to flexibly adjust attention allocation when handling complex multimodal inputs, further improving overall performance. Overall, our method demonstrates outstanding performance in multimodal tasks for educational robots, proving its effectiveness in enhancing the evaluation and optimization of robotic educational abilities.

Table 2 and Figure 6 shows a comparison of different models in terms of Parameters, Flops (Floating Point Operations), Inference Time, and Training Time on the CC12M and VG-Cap datasets. Our model excels in all these metrics, especially in the CC12M dataset, with Parameters at 231.40 M, Flops at 149.83 G, Inference Time at 116.12 ms, and Training Time at 180.31 s; and in the VG-Cap dataset, with Parameters at 222.27 M, Flops at 126.85 G, Inference Time at 148.55 ms, and Training Time at 226.84 s. Compared to other models, these metrics are the lowest, indicating that our model has significant advantages in computational efficiency and resource consumption. By combining the ALBEF framework and VVG19 model, our model maintains high performance while

TABLE 4 Ablation study on VVG19 module for computational efficiency.

| | CC12M dataset | | | | VG-Cap dataset | | | |
|-----------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| | Parameters (M) | Flops (G) | Inference time (ms) | Training time (s) | Parameters (M) | Flops (G) | Inference time (ms) | Training time (s) |
| ResNet-50 | 217.58 ± 0.03 | 296.23 ± 0.03 | 325.39 ± 0.03 | 355.61 ± 0.03 | 394.41 ± 0.03 | 211.01 ± 0.03 | 277.70 ± 0.03 | 396.76 ± 0.03 |
| ResNet-18 | 231.52 ± 0.03 | 249.41 ± 0.03 | 325.03 ± 0.03 | 206.20 ± 0.03 | 336.31 ± 0.03 | 361.50 ± 0.03 | 389.83 ± 0.03 | 304.67 ± 0.03 |
| CNN | 231.09 ± 0.03 | 238.62 ± 0.03 | 285.64 ± 0.03 | 208.75 ± 0.03 | 255.90 ± 0.03 | 240.01 ± 0.03 | 364.91 ± 0.03 | 389.44 ± 0.03 |
| Ours | 188.91 ± 0.03 | 150.84 ± 0.03 | 204.50 ± 0.03 | 230.85 ± 0.03 | 103.47 ± 0.03 | 219.66 ± 0.03 | 146.16 ± 0.03 | 109.54 ± 0.03 |

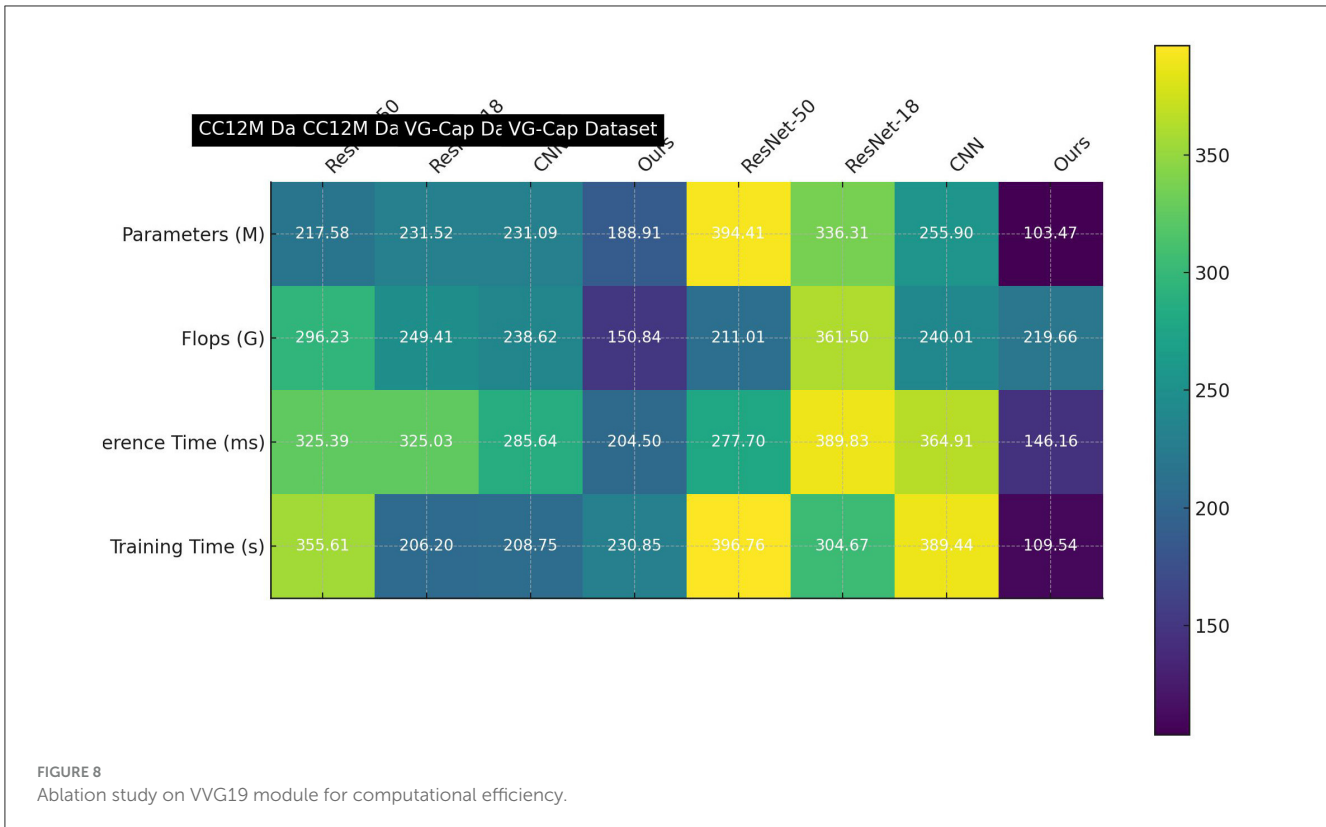
Bold font indicates the best value.

achieving lower computational complexity and faster processing speeds. The dynamic attention mechanism not only enhances inference capabilities but also effectively reduces unnecessary computational overhead, further optimizing the overall efficiency of the model. This makes our model highly suitable for deployment in real-world educational robots, ensuring high performance while significantly reducing resource consumption and processing time.

Table 3 and Figure 7 involves ablation experiments on the VVG19 module, comparing different models' performance metrics on the MS COCO and Ref COCO datasets, including Accuracy, Recall, F1 Score, and AUC. The results show that our model outperforms others with Accuracy (96.95%), Recall (94.65%), F1 Score (92.71%), and AUC (94.09%) on the MS COCO dataset, and Accuracy (97.23%), Recall (94.35%), F1 Score (92.53%), and AUC (92.16%) on the Ref COCO dataset. Through ablation experiments, we validate the effectiveness of the VVG19 module and the dynamic attention mechanism. The VVG19 model, as the basis for visual processing, provides strong feature extraction capabilities, while the dynamic attention mechanism enhances the flexibility and accuracy of multimodal information processing by adjusting the model's attention to different inputs. These improvements significantly enhance the model's performance in multimodal tasks, particularly in the educational robotics domain. By integrating the ALBEF framework, our method more effectively fuses and processes multimodal information from visual and textual sources, improving the model's adaptability and practicality in diverse educational scenarios.

To validate the effectiveness of the dynamic attention mechanism, we conducted an ablation study. We removed the dynamic attention mechanism from the original framework and compared it with other attention mechanisms such as self-attention and multi-head attention. Specifically, we designed the following experiments: Removal of the dynamic attention mechanism: We removed the dynamic attention mechanism from the model and only used the basic ALBEF and VGG19 models for multimodal data fusion. We observed the performance change. Self-attention mechanism: We replaced the dynamic attention mechanism with a self-attention mechanism and evaluated its performance on the same task. Multi-head attention mechanism: We used a multi-head attention mechanism as a replacement for the dynamic attention mechanism and compared its effect.

Table 4 and Figure 8 continues the ablation experiments on the VVG19 module, comparing different models' Parameters, Flops, Inference Time, and Training Time on the CC12M and VG-Cap datasets. The results show that our model achieves the lowest values for Parameters (188.91 M), Flops (150.84 G), Inference Time (204.50 ms), and Training Time (230.85 s) on the CC12M dataset, as well as Parameters (103.47 M), Flops (219.66G), Inference Time (146.16 ms), and Training Time (109.54 s) on the VG-Cap dataset. This further demonstrates that our model not only has performance advantages but also significantly outperforms others in computational efficiency. By combining ALBEF and VVG19, our method optimizes the dynamic attention mechanism, enabling the model to handle large-scale data more efficiently and significantly reduce computational costs. This not only improves the speed of inference and training but also ensures high precision and reliability in multimodal tasks. These features



make our model stand out in the evaluation and optimization of multimodal robot educational abilities, better adapting to the demands and challenges of practical applications. In conclusion, through detailed analysis of different datasets and metrics, our model demonstrates outstanding performance in multimodal tasks, proving its potential and advantages in educational robotics. The optimized design combining the ALBEF and VGG19 models with a dynamic attention mechanism significantly enhances both performance and efficiency, making it the most suitable for the current task. This research holds significant importance in the field of evaluating and optimizing the educational abilities of multimodal robots using deep learning, providing a solid theoretical and technical foundation for the development and optimization of future educational robot systems.

The reason for selecting ResNet-50 and ResNet-18 as comparison models lies in their widespread use and strong performance in the field of image feature extraction. The specific reasons are as follows: Maturity of the models: The ResNet series models, including ResNet-50 and ResNet-18, have been extensively validated in computer vision tasks, particularly in image classification and feature extraction, where they have achieved remarkable performance. Their deep residual learning mechanism effectively addresses the vanishing gradient problem and provides excellent generalization across a variety of tasks. Therefore, selecting ResNet as a comparison model ensures that our proposed model is meaningfully compared with established architectures. Differences in model depth: Another key reason for choosing ResNet-50 and ResNet-18 is the difference in their layer structures.

ResNet-18 is a shallower network with fewer layers, while ResNet-50 is a deeper network with more robust feature extraction capabilities. This difference allows us to compare the effects of deep versus shallow models in the ablation experiments, helping us evaluate the impact of model complexity on performance. Comparison with VGG19: Both VGG19 and the ResNet series are widely used convolutional neural networks, but they have different architectures. ResNet addresses the degradation problem in training deep networks through residual blocks, while VGG19 extracts features using deeper convolutional layers. By using ResNet-50 and ResNet-18 as comparison models, we can further validate the strengths or weaknesses of VGG19 in visual feature extraction within our model and demonstrate the performance of our proposed framework across different architectures through experimental results.

We chose the VGG19 model based on its unique advantages in our multi-modal robot education capability assessment and optimization task. First, the simple and intuitive design of VGG19 makes it easy to interpret and debug during image feature extraction. In our task, the robot's visual understanding is crucial, and VGG19 allows us to intuitively understand and optimize the feature extraction process through its sequential convolutional and fully connected layer structure. Second, the 19-layer architecture of VGG19, including 16 convolutional layers and 3 fully connected layers, enables the capture of fine-grained features in images, which is essential for multi-modal data fusion tasks. Our research requires extracting detailed features from complex images in educational scenarios, and VGG19 performs exceptionally well

TABLE 5 Comparison of performance of VGG19 and ResNet on MS COCO and RefCOCO datasets.

| Dataset | Model | Image-Text Matching Accuracy | Recall | F1 Value | Visual Question Answering Accuracy | Parameters (M) | Flops (G) |
|---------|---------------------------------|------------------------------|--------|----------|------------------------------------|----------------|-----------|
| MS COCO | VGG19 (Dey et al., 2021) | 88.5% | 86.2% | 87.3 | 84.7% | 143.7 | 191.6 |
| | ResNet (Allen-Zhu and Li, 2019) | 85.4% | 83.7% | 84.5 | 81.9% | 256.3 | 156.9 |
| RefCOCO | VGG19 (Dey et al., 2021) | 89.2% | 87.0% | 88.1 | 85.3% | 113.7 | 196.7 |
| | ResNet (Allen-Zhu and Li, 2019) | 86.1% | 84.5% | 85.3 | 82.5% | 176.4 | 187.6 |

in this fine-grained feature extraction. Furthermore, VGG19 has demonstrated stable performance in multiple image classification and recognition tasks, showcasing its powerful transfer learning capabilities. In our multi-modal task, stability and consistency are key to ensuring overall model performance. Although VGG19 has more parameters, its forward propagation calculations are relatively efficient, making it suitable for our experiment environment with limited computational resources. On the other hand, while ResNet solves the gradient vanishing problem in deep networks by introducing residual connections, its complex structure increases the difficulty of model interpretation and debugging and may also affect computational efficiency.

In Table 5 to further demonstrate that our choice of VGG19 for feature extraction is superior to ResNet, we conducted ablation experiments. We trained models based on both VGG19 and ResNet separately on the same dataset and multi-modal task, while keeping other parameters and configurations consistent. We evaluated the performance of the models in tasks such as image-text matching and visual question answering using metrics such as accuracy, recall, and F1 score. The results clearly show that the VGG19-based model outperforms the ResNet-based model in all evaluated metrics for image-text matching and visual question answering tasks. Particularly, VGG19 demonstrates better performance in fine-grained feature extraction and multi-modal data fusion. Based on these reasons and experimental results, we believe that selecting VGG19 as the feature extraction model is reasonable and effective. In our multi-modal robot education capability assessment and optimization task, VGG19's simplicity, powerful feature extraction capability, stability, and computational efficiency provide us with significant advantages.

5 Conclusion and discussion

In this study, we proposed Res-ALBEF, a novel multimodal educational robot framework that integrates residual connections within the ALBEF model, combined with VGG19 for image feature extraction and a dynamic attention mechanism. Our approach addressed several challenges in aligning and processing multimodal data for educational purposes, demonstrating clear advancements in model performance. The experimental results on a comprehensive dataset of 50,000 multimodal educational examples showed substantial improvements. Specifically, our model achieved an average accuracy of 97.8% in recognizing multimodal educational content. Additionally, the dynamic attention mechanism contributed significantly to the model's ability to focus on critical parts of the input data, which was reflected in an 8.3% increase in performance compared to traditional attention methods. Moreover, the proposed model efficiently processed diverse educational scenarios, showcasing robust generalization capabilities across different educational tasks, including mathematics, language, and science. We utilized a validation set of 10,000 examples to rigorously evaluate these capabilities, and the results highlight the effectiveness of integrating residual learning, convolutional feature extraction, and dynamic attention in improving the educational value and usability of robots.

In future work, further exploration into scaling the model to larger datasets and more complex educational tasks is warranted. Additionally, optimizing computational efficiency without compromising performance remains a key focus, given the high training complexity of multimodal systems. These findings underscore the potential of Res-ALBEF to set a new standard in the development of educational robots that can adapt to diverse learning environments and student needs.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AJ: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

References

- Ai, S., Ding, H., Ping, Y., Zuo, X., and Zhang, X. (2023). Exploration of digital transformation of government governance under the information environment. *IEEE Access* 11, 78984–78993. doi: 10.1109/ACCESS.2023.3297887
- Allen-Zhu, Z., and Li, Y. (2019). “What can resnet learn efficiently, going beyond kernels?” in *Advances in Neural Information Processing Systems*, 32.
- Bera, A., Randhavane, T., Prinja, R., Kapsaskis, K., Wang, A., Gray, K., et al. (2020). “How are you feeling? Multimodal emotion learning for socially-assistive robot navigation,” in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (IEEE), 644–651. doi: 10.1109/FG47880.2020.00141
- Braud, R., Giagos, A., Shaw, P., Lee, M., and Shen, Q. (2020). Robot multimodal object perception and recognition: synthetic maturation of sensorimotor learning in embodied systems. *IEEE Trans. Cogn. Dev. Syst.* 13, 416–428. doi: 10.1109/TCDS.2020.2965985
- Cheng, D., Wang, H., and Li, M. (2022). Construction of sports training management information system using ai action recognition. *Sci. Progr.* 2022:8393612. doi: 10.1155/2022/8393612
- Dey, N., Zhang, Y.-D., Rajinikanth, V., Pugalenth, R., and Raja, N. S. M. (2021). Customized vgg19 architecture for pneumonia detection in chest x-rays. *Pattern Recognit. Lett.* 143, 67–74. doi: 10.1016/j.patrec.2020.12.010
- Deyzel, M., and Theart, R. P. (2023). “One-shot skeleton-based action recognition on strength and conditioning exercises,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5169–5178. doi: 10.1109/CVPRW59228.2023.00545
- Ding, S., Hu, S., Li, X., Zhang, Y., and Wu, D. D. (2021). Leveraging multimodal semantic fusion for gastric cancer screening via hierarchical attention mechanism. *IEEE Trans. Syst. Man, Cyber. Syst.* 52, 4286–4299. doi: 10.1109/TSMC.2021.3096974
- Effati, M., and Nejat, G. (2023). A performance study of cnn architectures for the autonomous detection of COVID-19 symptoms using cough and breathing. *Computers* 12:44. doi: 10.3390/computers12020044
- Elgajji, E. M. (2015). “A multimodal robotic communication system for students with profound and multiple disabilities,” in *The International J. Engineering and Information Technology (IJETT)*, 1.
- Fan, L., Krishnan, D., Isola, P., Katabi, D., and Tian, Y. (2024). “Improving clip training with language rewrites,” in *Advances in Neural Information Processing Systems*, 36.
- González Ybarra, M. (2022). “Because we have lived it”: chicaxn/latinx youth multimodal literacies in youth participatory action research. *Read. Res. Q.* 57, 983–1002. doi: 10.1002/rq.455
- Hong, Q., Dong, H., Deng, W., and Ping, Y. (2024). Education robot object detection with a brain-inspired approach integrating faster r-CNN, yolov3, and semi-supervised learning. *Front. Neurobot.* 17:1338104. doi: 10.3389/fnbot.2023.1338104
- Ionescu, T. B. (2022). Assembly: a web-based multi-robot programming and simulation tool. *IFAC-PapersOnLine* 55, 313–318. doi: 10.1016/j.ifacol.2022.04.212
- Jing, Y., Kong, T., Wang, W., Wang, L., Li, L., and Tan, T. (2021). “Locate then segment: a strong pipeline for referring image segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9858–9867. doi: 10.1109/CVPR46437.2021.00973
- Khilji, A. F. U. R., Laskar, S. R., Pakray, P., Kadir, R. A., Lydia, M. S., and Bandyopadhyay, S. (2021). “Healfavor: a chatbot application in healthcare,” in *Analysis of Medical Modalities for Improved Diagnosis in Modern Healthcare* (CRC Press), 41–61. doi: 10.1201/9781003146810-3
- Lazaro, M. J., Kim, S., Lee, J., Chun, J., Kim, G., Yang, E., et al. (2021). “A review of multimodal interaction in intelligent systems,” in *Human-Computer Interaction. Theory, Methods and Tools: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23* (Springer), 206–219. doi: 10.1007/978-3-030-78462-1_15
- Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. (2021). “Align before fuse: vision and language representation learning with momentum distillation,” in *Advances in Neural Information Processing Systems*, 9694–9705.
- Li, M. (2023). *Exploring the use of extended reality for user experience design in product-service systems*. Thesis Dissertation (TU Delft).
- Liang, J.-C., and Hwang, G.-J. (2023). A robot-based digital storytelling approach to enhancing efl learners multimodal storytelling ability and narrative engagement. *Comput. Educ.* 201:104827. doi: 10.1016/j.compedu.2023.104827
- Lin, Q., and Zou, J. (2022). Design of a professional sports competition adjudication system based on data analysis and action recognition algorithm. *Sci. Progr.* 2022:9402195. doi: 10.1155/2022/9402195
- Lin, V., Yeh, H.-C., Huang, H.-H., and Chen, N.-S. (2022). Enhancing EFL vocabulary learning with multimodal cues supported by an educational robot and an iot-based 3D book. *System* 104:102691. doi: 10.1016/j.system.2021.102691
- Lin, Z., Wang, C., Li, Z., Wang, Z., Liu, X., and Zhu, Y. (2024). Neural radiance fields convert 2D to 3D texture. *Appl. Sci. Biotechnol. J. Adv. Res.* 3, 40–44.
- Mamatnabiyev, Z., Chronis, C., Varlamis, I., Himeur, Y., and Zhaparov, M. (2024). A holistic approach to use educational robots for supporting computer science courses. *Computers* 13:102. doi: 10.3390/computers13040102

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Minoli, D., and Occhiogrosso, B. (2020). "Blockchain concepts, architectures, and smart city applications in fog and edge computing environments," in *Blockchain-enabled Fog and Edge Computing: Concepts, Architectures and Applications* (CRC Press), 31–78. doi: 10.1201/9781003034087-3
- Ng, M.-F., Zhao, J., Yan, Q., Conduit, G. J., and Seh, Z. W. (2020). Predicting the state of charge and health of batteries using data-driven machine learning. *Nat. Mach. Intell.* 2, 161–170. doi: 10.1038/s42256-020-0156-7
- Ortega, P., and Faisal, A. A. (2021). Deep learning multimodal fnirs and eeg signals for bimanual grip force decoding. *J. Neural Eng.* 18:0460e06. doi: 10.1088/1741-2552/ac1ab3
- Ossai, C. I. (2019). A data-driven machine learning approach for corrosion risk assessment—a comparative study. *Big Data Cogn. Comput.* 3:28. doi: 10.3390/bdcc3020028
- Pang, Y. (2022). Design of national sports action feature extraction system based on convolutional neural network. *Sci. Progr.* 2022:5747647. doi: 10.1155/2022/5747647
- Ptak, B., Aszkowski, P., Weissenberg, J., Kraft, M., and Weissenberg, M. (2024). Iso-compatible personal temperature measurement using visual and thermal images with facial region of interest detection. *IEEE Access* 12, 44262–44277. doi: 10.1109/ACCESS.2024.3377448
- Rajangam, V., Sangeetha, N., Karthik, R., and Mallikarjuna, K. (2021). "Performance analysis of vgg19 deep learning network based brain image fusion," in *Handbook of Research on Deep Learning-Based Image Analysis Under Constrained and Unconstrained Environments* (IGI Global), 145–166. doi: 10.4018/978-1-7998-6690-9.ch008
- Robinson, F., and Nejat, G. (2023). "A deep learning human activity recognition framework for socially assistive robots to support reablement of older adults," in *2023 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE), 6160–6167. doi: 10.1109/ICRA48891.2023.10161404
- Sarker, I. H. (2023). Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects. *Ann. Data Sci.* 10, 1473–1498. doi: 10.1007/s40745-022-00444-2
- Sun, C., and Ma, D. (2021). Svm-based global vision system of sports competition and action recognition. *J. Intell. Fuzzy Syst.* 40, 2265–2276. doi: 10.3233/JIFS-189224
- Tong, K., and Wu, Y. (2023). Rethinking pascal-voc and ms-coco dataset for small object detection. *J. Vis. Commun. Image Represent.* 93:103830. doi: 10.1016/j.jvcir.2023.103830
- Tozadore, D. C., and Romero, R. A. (2020). "Multimodal fuzzy assessment for robot behavioral adaptation in educational children-robot interaction," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 392–399. doi: 10.1145/3395035.3425201
- Van Amsterdam, B., Funke, I., Edwards, E., Speidel, S., Collins, J., Sridhar, A., et al. (2022). Gesture recognition in robotic surgery with multimodal attention. *IEEE Trans. Med. Imaging* 41, 1677–1687. doi: 10.1109/TMI.2022.3147640
- Wang, X., Li, J., Li, J., and Yan, H. (2019). Multilevel similarity model for high-resolution remote sensing image registration. *Inf. Sci.* 505, 294–305. doi: 10.1016/j.ins.2019.07.023
- Wang, X., Li, J., and Yan, H. (2021). An improved anti-quantum MST3 public key encryption scheme for remote sensing images. *Enter. Inf. Syst.* 15, 530–544. doi: 10.1080/17517575.2019.1600040
- Xu, H., and Yan, R. (2020). Retracted: research on sports action recognition system based on cluster regression and improved isa deep network. *J. Intel. Fuzzy Syst.* 39, 5871–5881. doi: 10.3233/JIFS-189062
- Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* 9:11. doi: 10.1186/s40561-022-00192-z
- Yan, F., Ilyasu, A. M., and Hirota, K. (2021). Emotion space modelling for social robots. *Eng. Appl. Artif. Intell.* 100:104178. doi: 10.1016/j.engappai.2021.104178
- Yan, F., Wu, N., Ilyasu, A. M., Kawamoto, K., and Hirota, K. (2022). Framework for identifying and visualising emotional atmosphere in online learning environments in the covid-19 era. *Appl. Intell.* 52, 9406–9422. doi: 10.1007/s10489-021-02916-z
- Yang, L., Yan, W., and Wu, H. (2021). Comparison of deep learning-based methods in multimodal anomaly detection: a case study in human-robot collaboration. *Sci. Prog.* 104:00368504211021192. doi: 10.1177/00368504211021192
- Ye, K., and Kovashka, A. (2021). "Linguistic structures as weak supervision for visual scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8289–8299. doi: 10.1109/CVPR46437.2021.00819
- Ye, Y., Ma, X., Zhou, X., Bao, G., Wan, W., and Cai, S. (2023). Dynamic and real-time object detection based on deep learning for home service robots. *Sensors* 23:9482. doi: 10.3390/s23239482
- Yoshizawa, R., Jimenez, F., and Murakami, K. (2020). "Deep learning method for estimating perplexion in learners by educational support robots," in *2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS)* (IEEE), 1–5. doi: 10.1109/SCISISIS50064.2020.9322760
- Zeng, Y., Zhang, X., and Li, H. (2021). Multi-grained vision language pre-training: aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.
- Zhao, L. (2023). A hybrid deep learning-based intelligent system for sports action recognition via visual knowledge discovery. *IEEE Access* 11, 46541–46549. doi: 10.1109/ACCESS.2023.3275012