Multimodal Fusion Comprehension Model for English Listening and Literature Appreciation Based on the Perceiver IO Architecture

Jing Li
Department of Public Foreign Language Teaching, Luohe Medical College, Luohe 462000, Henan, China
Corresponding author: Jing Li (e-mail: 15639569021@163.com)

ABSTRACT Traditional English multimodal listening comprehension models suffer from semantic fragmentation and weak task transferability because they lack a unified cross-modal attention mechanism, have difficulty processing high-dimensional images and long-duration speech signals simultaneously, and often rely on loosely coupled fusion strategies such as post -concatenation or simple weighting. To address these issues, this paper uses Wav2Vec 2.0 and BERT to achieve semantic alignment between speech and text, ensuring spatiotemporal consistency of linguistic information. A unified input representation structure is then constructed to improve representational consistency across modalities. A fusion encoding module centered on Perceiver IO is built, and a cross-attention mechanism is introduced to realize contextual coupling among modalities. Finally, a multi-objective prediction structure is implemented at the output layer to identify rhetorical elements, sentiment nodes, and key sentences in literary listening materials, including novels, poetry recitations, and technical literary expressions. The model achieves sentiment classification accuracy of 0.86 with F1 of 0.85, and F1 above 0.80 in cross-text rhetorical recognition. The speech-text alignment distance decreases to nearly 0.2, with BLEU ranging from 0.78 to 0.85, confirming effective multimodal fusion and comprehension.

INDEX TERMS multimodal understanding, english listening, perceiver io, semantic alignment, cross-modal attention

I. INTRODUCTION

As foreign language teaching reforms continue to advance, multimodal intelligent technology has emerged as a key research direction in language learning, particularly for acquiring the specialized vocabulary and communication skills necessary for the industry. In English listening instruction, the traditional information reception model, centered around a single audio source, has gradually shifted to a comprehensive comprehension task that integrates multiple sources of information, including text, audio, and video [1,2]. Literary listening materials, as a typical carrier of cultural aesthetics and deep context in English teaching, contain multiple expressions of emotional intonation, complex rhetorical structures, and cultural connotations, placing higher demands on learners' information processing abilities and cognitive structures [3,4]. The introduction of artificial intelligence (AI) technology is gradually making language learning more intelligent and systematic. For the multimodal, complex cognitive context of literary listening, integrating heterogeneous data sources such as audio, text, and visuals through deep learning models to achieve a deep understanding of semantics, context, and emotion has become a key challenge in current intelligent teaching research [5,6]. At the same time, multimodal models generally lack the ability to discern emotional expression and rhetorical intent in listening materials, making it difficult to effectively assist learners in constructing an aesthetic understanding of literary content. Traditional single-speech listening models are not very accurate in understanding literary materials. Literary listening materials contain an average of 3.8 rhetorical devices, 2.1 emotional level transitions, and 1.5 cultural metaphors per thousand words, representing a cognitive complexity higher than that of conventional listening materials, significantly increasing learners' information processing load.

Existing literature has extensively explored the application of multimodal learning in language teaching. Research focuses on optimizing English speech recognition accuracy, synchronizing video subtitles, extracting text semantics, and integrating feature representations [7,8]. In audio understanding, researchers generally use convolutional neural networks, CTC (Connectionist Temporal Classification)

decoding mechanisms, and end-to-end speech recognition models to achieve speech recognition and transcription, and attempt to utilize emotion recognition networks to enhance intonation-level information processing [9,10]. Traditional multimodal model fusion approaches are often based on rule-driven or branch-based integration, failing to establish a universal structure for unified input and output, resulting in poor transferability in practical applications [11,12]. These shortcomings in current research have led to prominent problems with multimodal models in handling high-level English literary listening tasks, such as semantic fragmentation, feature redundancy, and superficial understanding [13,14]. To enhance multimodal literary comprehension, this study proposes a novel architecture that integrates the inherent connections between sentiment recognition and rhetorical analysis and designs a hierarchical interaction mechanism. Compared to traditional models, the Perceiver IO architecture dynamically achieves cross-modal temporal alignment through a scalable cross-attention mechanism, effectively reducing the alignment error between speech and text without the need for pre-set alignment rules.

To address these challenges, this paper proposes a multimodal model based on the Perceiver IO (Perceiver Input-Output) architecture specifically for literary appreciation and comprehension in English listening. As a structured neural network with universal input and output processing capabilities, Perceiver IO overcomes the input format and dimensionality limitations faced by traditional Transformers when processing multimodal data. Leveraging a scalable cross-attention mechanism, it can accept heterogeneous input data and generate predicted outputs of arbitrary structure. Independent of the design details of a specific modality, this architecture demonstrates high flexibility and generalization potential, making it particularly suitable for multimodal fusion tasks involving complex semantic hierarchies [15,16]. In this study, this architecture was applied to uniformly encode the following information: audio signals (encoded via Wav2Vec 2.0), text information encoded via BERT (Bidirectional Encoder Representations from Transformers), and visual information (features extracted by an image encoder). A cross-modal attention module enabled deep interaction and fusion between different modal information. At the output, the model combined sequence labeling and classification tasks to automatically identify literary elements and acquire contextual awareness [17,18]. This study, based on real-world teaching scenarios, combined audio, subtitled text, and visual materials to enhance the comprehension of literary listening content through multi-source data modeling. The method design included: constructing a speech-text alignment mechanism to enhance temporal synchronization perception; designing a unified input structure to enhance modal consistency; utilizing Perceiver IO and a cross-attention mechanism for modal fusion; implementing multi-target prediction at the output layer to identify rhetorical elements, emotional nodes, and key sentences; and finally, incorporating a teaching-oriented visual explanation module. This teaching visualization module innovatively utilizes the WAAS (Weighted Attention Aggregation Score) scoring system. Through this approach, the model can achieve task-oriented listening context modeling, structured analysis of literary content, and identification and expression of aesthetic elements in teaching applications, driving the evolution of intelligent language understanding systems towards deeper cognitive support. On a theoretical level, this research breaks through the limitations of modal input formats and constructs a unified representation framework for literary listening comprehension. On an applied level, the model achieves a macro F1 score of 0.80 for literary element recognition in real-world teaching scenarios, and its teaching applicability is verified through expert evaluation. This offers a transferable technical path and a practical paradigm for implementing multimodal intelligent teaching systems, which can be widely applied to domains like engineering and design training.

II. CONSTRUCTING ENGLISH LITERARY COMPREHENSION WITH FUSION PERCEPTION

A. CONSTRUCTING A SEMANTIC ALIGNMENT MECHANISM FOR TEXT AND SPEECH CONTENT

1) Audio Semantic Feature Extraction

This study first extracts semantic features from audio data from English literature listening comprehension. This study uses a pre-trained Wav2Vec2.0 model, which, based on a self-supervised learning framework, can directly learn effective speech representations from raw audio waveforms [19,20]. The audio input is sampled as a 16kHz mono signal and segmented into 20ms frames. The convolutional encoder layer of Wav2Vec2.0 extracts features from each frame of audio, generating a high-dimensional hidden representation. Subsequently, based on the Transformer encoder structure, the model encodes local frame features into a sequence of information vectors with contextual semantics through contextual relationship modeling. To ensure the model's applicability in the context of English literature listening, this paper fine-tuned Wav2Vec 2.0 using a speech corpus containing a variety of literary readings. During training, the Adam optimizer was used with a learning rate of 1e-4, a batch size of 32, and 50 epochs. Ultimately, the model's ability to capture speech semantics in specific audio styles was improved.

2) Text Semantic Encoding and Alignment Strategy

The text portion uses the BERT-Base model for semantic encoding. The text input is first segmented using WordPiece, and the sequence length is limited to 128 tokens. The BERT encoder outputs a 768-dimensional context embedding vector corresponding to each token. To achieve audio and text alignment, this paper designs a dual-channel alignment mechanism. This mechanism first roughly maps the time range of the audio timestamps and the sentence boundary information of the text subtitles. The Dynamic Time Warping (DTW) algorithm then refines the correspondence

between audio frames and text tokens. DTW uses the cosine distance between audio frame features and text token embeddings as a cost function to find the optimal alignment path, ensuring temporal synchronization while minimizing semantic differences [21,22]. After alignment, the model refines the semantic fusion of audio and text using a cross-modal attention module. Specifically, the model uses a text embedding sequence as the query and audio features as the key and value. A multi-head attention mechanism is used to calculate the influence weights of the audio on the text, generating a joint semantic representation. The model uses 8 attention heads, with a single head dimension of 64. Residual connections and LayerNorm ensure training stability. This mechanism effectively mitigates semantic drift and alignment errors between modalities, improving the accuracy and robustness of the representation.

Figure 1 illustrates the semantic alignment mechanism between speech and text. The left panel shows the cosine similarity matrix between audio frames and text tokens. The horizontal axis represents the text token index, and the vertical axis represents the audio frame index. The color depth reflects the degree of similarity between the semantic features of the two, showing the semantic matching relationship between audio and text at different time points. The right image in Figure 1 shows the DTW alignment path. The background is the distance matrix between audio and text features. Color represents the distance, with darker areas corresponding to smaller distances and more similar features. The red line represents the optimal alignment path found by DTW, and the green and red dots mark the starting and ending points of the path, respectively, illustrating the temporal alignment between text tokens and audio frames.

3) Speech-to-Text Alignment Formula Description

Let $A = \{a_1, a_2, \ldots, a_m\}$ denote the audio frame feature sequence and $T = \{t_1, t_2, \ldots, t_n\}$ denote the text token embedding sequence. Dynamic Time Warping defines the cost matrices $D \in R^{m \times n}$ and $D(i, j)$ as satisfying Formula 1:

$$D(i, j) = 1 - \cos(a_i, t_j) \tag{1}$$

$\cos(a_i, t_j)$ is the cosine similarity between feature vectors. The optimal alignment path $P = \{(i_k, j_k)\}$ satisfies Formula 2:

$$P = \arg\min_{P} \sum_{k} D(i_k, j_k) \tag{2}$$

The path constraint ensures that the temporal order is monotonically non-decreasing. Subsequently, the alignment weight is calculated through the cross-modal multi-head attention function $Attn(Q, K, V)$ as shown in Formula 3:

$$Attn(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{3}$$

$Q = TW_Q$, $K = AW_K$, $V = AW_V$, W_Q, W_K, W_V are linear projection matrices, and d_k is the key dimension. This process completes the semantically weighted mapping of audio to text, generating a fused, aligned representation.

The reason for introducing Wav2Vec2.0 and BERT for pre-alignment before Perceiver IO is the significant semantic gap between the original audio waveform and the literal text: the former contains paralinguistic information such as prosody and pauses, while the latter carries syntactic and rhetorical structures. If the original signal is directly input into Perceiver, its cross-attention mechanism needs to simultaneously complete low-level feature extraction and high-level semantic alignment, which can easily lead to optimization difficulties with limited training data. Through pre-alignment, we map the audio into contextualized frame embeddings and encode the text into deep token embeddings. On this basis, Perceiver IO's cross-attention mechanism completes cross-modal semantic fine-tuning and global fusion, thus forming a two-stage strategy of "coarse alignment - fine fusion". This design is not redundant, but rather a hierarchical decoupling of modal heterogeneity and semantic complexity.

B. DESIGN OF A UNIFIED MULTIMODAL INPUT REPRESENTATION

1) Heterogeneous Modal Feature Encoding

Before achieving multimodal fusion, the three heterogeneous data types of speech, text, and vision must be converted into a vector sequence in a unified format. For audio features, this study continues the output of the Wav2Vec2.0 model described above, selecting the hidden vector of the final Transformer layer as the audio semantic representation [23,24]. The audio vector length is fixed to 256 time steps, with a feature dimension of 768 per step. Linear interpolation or truncation is used to unify the length to ensure input sequence normalization.

For text features, the token-level embedding output by the BERT model is inherited, with the text sequence length limited to 128 tokens, and the embedding dimension per token is also 768. The text embedding uses positional encoding and segmented encoding to enhance spatial information. Positional encoding uses a trainable absolute position vector with the same dimensionality as the embedding to preserve the sequential nature of the text.

For the image modality, this paper employs the Vision Transformer (ViT-Base/16) as the visual encoder. The input image is cropped to 224×224 pixels and divided into 14×14 image patches. Each patch is linearly projected and then superimposed with learnable positional codes. A 12-layer Transformer encoder outputs a 768-dimensional global feature vector. Finally, all patch embeddings except the [CLS] token are taken as the visual sequence (196 tokens in total) and unified to 768 dimensions through linear projection to match the audio and text dimensions. This sequence is downsampled to 64 tokens and then concatenated with the audio and text sequences to form a unified input.

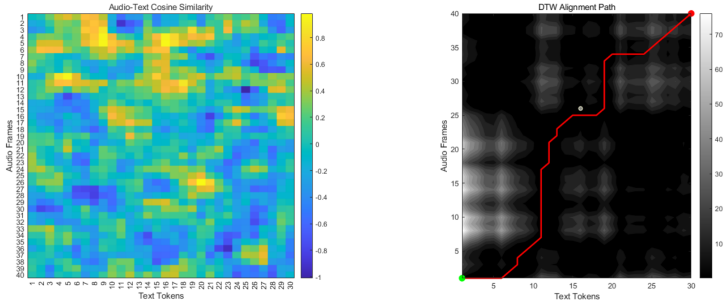

FIGURE 1. Semantic Alignment Mechanism between Speech and Text

2) Constructing a Unified Modality Representation

The feature vector sequences for the three modalities are: audio sequence $A \in R^{256\times768}$, text sequence $T \in R^{128\times768}$, and visual sequence $V \in R^{64\times768}$. These three sequences are concatenated along the time dimension to form the overall input sequence $X \in R^{448\times768}$, as shown in Formula 4:

$$X = Concat(A, T, V) \tag{4}$$

To clearly distinguish between different modal information, this paper introduces a modal encoding vector M, which is added to the input sequence token by token. This modal encoding vector is a trainable parameter with a dimension of 448×768. It is segmented by modality, ensuring that the Perceiver IO model can accurately identify and process the features of each modality. A unified sequence X+M serves as the input token sequence for Perceiver IO. The sequence length of 448 maintains a relatively compact size, balancing information integrity and model computational efficiency. The input undergoes a normalization layer to reduce scale differences between modalities.

In addition, to further improve the model's sensitivity to modal interactions, this paper proposes an input perturbation strategy. During training, a small Gaussian noise perturbation with a mean of 0 and a standard deviation of 0.02 is applied to the audio, text, and visual tokens in the input sequence, enhancing the model's generalization and robustness.

Figure 2 shows three key visualizations of the unified representation of multimodal features. The violin plot on the left compares the distribution of audio, text, and visual features. The Y-axis represents the normalized feature value, showing that audio features are concentrated around 0.5, text features are concentrated around -0.8, and visual features are distributed towards 1.5, reflecting the significant differences in the statistical properties of the original modalities. The PCA (Principal Component Analysis) dimensionality reduction plot in the middle maps the high-dimensional fused features into a two-dimensional space. The X-axis and Y-axis represent principal component 1 (68.3% variance explained) and principal component 2 (21.7% variance explained), respectively. Features from different modalities exhibit good clustering in low-dimensional space, demonstrating that the model successfully preserves modal diversity and structure within a unified representation layer. The t-SNE (t-Distributed Stochastic Neighbor Embedding) plot on the right further reveals the distribution structure of the fused features. Each modality is clearly distributed in the two-dimensional space after nonlinear dimensionality reduction, verifying that the modal encoding and fusion strategy strikes a balance between maintaining the independence and uniformity of modal features.

3) Perceiver IO Input Mapping Optimization

Perceiver IO is the core architecture of this study. Its input layer needs to process variable-length, heterogeneous, and multimodal sequences. To this end, an input mapping function finput is designed to effectively map high-dimensional inputs to low-dimensional latent space, as shown in Formula 5:

$$Z = f_{input}(X + M + P) \tag{5}$$

$Z \in R^{L\times D}$ represents the latent space representation, L=512 represents the number of latent tokens, and D=1024 represents the latent dimension. The input mapping utilizes a learnable fusion of linear projection and positional encoding, with the mapping weights optimized via backpropagation. This design ensures Perceiver IO's ability to compress multimodal information while addressing the traditional Transformer's excessive computational resource consumption when processing long sequences, improving the model's efficiency and performance in multimodal English listening comprehension tasks.

C. LITERARY CONTEXT-AWARE FUSION ENCODER MODULE

1) Construction of a Cross-Modal Contextual Fusion Mechanism

This study builds a cross-modal fusion architecture based on a unified input representation, centered around Perceiver IO, to achieve contextual semantic coupling of audio, text, and visual information. The key feature of the Perceiver IO model is that its input can be a sequence of tokens of any structure. Internally, a fixed number of latent arrays interact with the input in multiple rounds, capturing the global

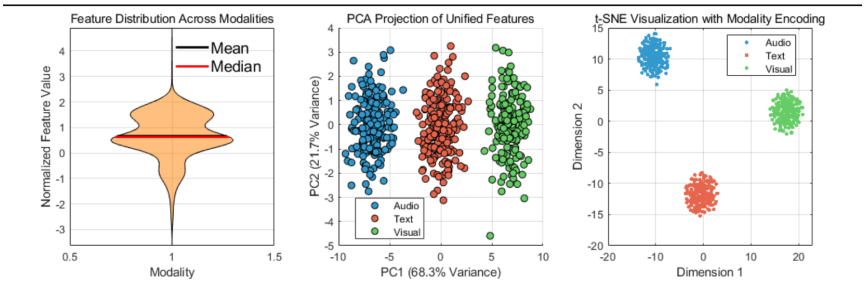

FIGURE 2. Visualization of the Unified Representation of Multimodal Features

context and potential relationships between modalities. This paper sets the number of latent tokens to 512 and the dimension to 1024, using a four-layer stacked cross-attention and self-attention module combination to accomplish the semantic fusion task.

The input sequence is the multimodal fusion representation $X \in R^{448\times768}$ constructed in the previous section, consisting of sequentially concatenated audio, text, and visual tokens. This sequence first interacts with the latent array $Z \in R^{512\times1024}$ via the cross-attention module. The attention mechanism uses the latent token as the query vector and performs a weighted aggregation operation on all tokens in the input sequence, thereby encoding the significant features of the heterogeneous modalities into the latent representation. The calculation method of this process is as shown in Formula 6:

$$CrossAttn(Q,K,V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (6)$$

The query matrix $Q = ZW_Q$, key matrix $K = XW_K$, value matrix $V = XW_V$, and weight matrix $W_Q, W_K, W_V \in R^{1024\times1024}$ are training parameters for $d_k = 1024$. This cross-attention module ensures that Perceiver IO dynamically focuses on different modalities based on the target of each latent token, effectively alleviating the problem of uneven representation between modalities and providing deep interactive support for downstream semantic modeling.

2) Contextual Hierarchy Modeling and Support for Rhetorical Feature Recognition

Literary listening materials contain significant contextual hierarchical structures, such as syntactic and semantic constructions, emotional fluctuations and rhythmic changes, and symbolic semantics and cultural references. To enhance the model's ability to perceive these structures, this paper introduces a multi-task perception mechanism at the output stage of the Perceiver IO and hierarchically models the contextual dimensions of the latent representation. Specifically, the output latent sequence is projected onto multiple task subspaces: sentence-level rhetorical figure identification, paragraph sentiment classification, and contextual turning point location. A separate loss function is defined for each task, employing a weighted joint optimization strategy.

D. MULTI-LAYER SEMANTIC OUTPUT AND IDENTIFICATION OF APPRECIATION ELEMENTS

This paper operationalizes "literary appreciation" into three quantifiable and teachable core sub-tasks: (1) rhetorical element identification : capturing the language techniques used by the author (such as metaphor and personification) to reflect the aesthetic structure of the text; (2) emotional node location : identifying the points of emotional polarity shift to reflect the narrative tension and psychological changes of the characters; (3) key sentence extraction : selecting sentences with high information density or strong thematic content to support the overall interpretation of the text. These three together constitute the basis for the three-dimensional understanding of literary texts in terms of "form-emotion-meaning", transforming the abstract "appreciation" into a trainable and assessable multi-objective prediction task.

1) Text Sequence Annotation Mechanism and BIO Structure Modeling

The primary goal of this module is to extract key semantic units of literary value from the fused representation. Through a multi-task mechanism, it automatically recognizes literary elements such as symbolic imagery, rhetorical structure, and emotional attitude. First, the latent token sequence $Z \in R^{512\times1024}$ output by the Perceiver IO fusion module is reduced to a task-matched dimension through a remapping projection layer. This is then used for decoding in the sequence labeling and classification tasks, respectively. For sentence-level or word-level text annotation, a serialized decoding strategy is employed, in which each text token is associated with a label, and the BIO tagging system (Begin, Inside, Outside) is used to model semantic segment boundaries. Let the input token sequence be $T = \{t_1, t_2, \ldots, t_n\}$, and its corresponding output be $H = \{h_1, h_2, \ldots, h_n\} \in R^{n\times d}$. Each h_i is mapped to the category space via linear projection and input into the Conditional Random Field (CRF) layer for label decoding. The CRF layer considers the transition probability between labels to ensure that the output sequence has structural consistency, avoiding confusion of label boundaries or illegal structures. The CRF maximization objective is Formula 7:

$$L_{CRF} = -\log \frac{e^{s(x,y)}}{\sum_{y'\in Y_x} e^{s(x,y')}} \quad (7)$$

s(x,y) is the scoring function for a given input sequence x and label sequence y, and Yx is the set of all possible label sequences. This mechanism enhances structural modeling in semantically dense text and is suitable for identifying literary elements that require continuous annotation, such as "symbolic imagery," "lyrical language," and "structural transitions."

2) Multi-Task Sentiment and Sentence Metaclassification Branch Design

In addition to the structural annotation task, the model constructs two semantic classification branches in parallel to assist in analyzing the overall sentiment and attitude of the sentence meta-level and the focus of literary expression [25,26]. The emotion recognition subtask uses sentence-level tokens from the fused representation to construct sentence vectors. After average pooling, these vectors are fed into a single-layer fully connected network for three-category output, with the label categories being positive emotion, neutral description, and negative emotion. The model optimization objective uses the standard cross-entropy loss, as shown in Formula 8:

$$L_{emotion} = -\sum_{i=1}^{C} y_i \log \hat{y}_i \tag{8}$$

C=3 is the number of sentiment categories, yi is the one-hot vector of the true label, and $\hat{y}$i is the predicted probability distribution after softmax. This module helps model the emotional evolution cues in literary contexts and provides pragmatic support for subsequent sentence-level semantic interpretation. The key sentence identification task aims to identify text units with high information density or emotional focus. Unlike sequence labeling, its goal is to classify sentence units into two categories (key/non-key). It adopts a sigmoid output structure based on sentence vectors and uses binary cross-entropy as the loss function. The sentence embeddings for this part are also derived from the text token sequence output from the Perceiver IO. Sentence representations are obtained through mean pooling, and classification predictions are performed directly. To improve the model's sensitivity to long-range syntactic and semantic dependencies, this branch classifier introduces a multi-head attention residual structure, performing contextual enhancement before constructing the sentence embeddings. A four-head attention structure is used, and the embedding dimension is maintained at 512.

3) Multi-Objective Joint Optimization and Semantic Output Strategy

The overall model adopts a multi-objective joint training strategy, integrating the CRF sequence labeling loss, sentiment classification loss, and key sentence recognition loss. The combined loss function is defined as Formula 9:

$$L_{total} = \lambda_1 L_{CRF} + \lambda_2 L_{emotion} + \lambda_3 L_{sentence} \tag{9}$$

$\lambda_1, \lambda_2, \lambda_3$ are the task weight parameters. This paper adjusts the values to 0.5, 0.3, and 0.2 based on the performance of the development set to ensure that the structural labeling task dominates the training. All submodules share the pre-fusion representation, and gradients are jointly back-propagated during parameter updates. To prevent gradient skew caused by label imbalance, the model introduces label balancing weights during training. Focal Loss is used to enhance difficult samples for the sentiment and sentence meta-classification tasks, with a focal loss coefficient of 2.

The left panel of Figure 3 shows the changing trends in the loss values for the three subtasks and the overall F1 score improvement of the model during multi-task training. The horizontal axis represents the number of training epochs, the left vertical axis represents the loss value, and the right vertical axis represents the F1 score. As training progresses, the CRF loss, sentiment recognition loss, and key sentence recognition loss all gradually decrease and converge, indicating that the model achieves good optimization results across various tasks. At the same time, the F1 score showed a clear upward trend, ultimately stabilizing above 0.6, demonstrating that the fusion training strategy effectively improved overall performance. Figure 3 (right) shows the precision-recall relationship curves for three literary elements: metaphor, symbol, and emotion, with recall on the horizontal axis and precision on the vertical axis. All three elements maintained high precision within a high recall range, with the emotion category achieving the highest AP (Average Precision) value of 0.86, demonstrating the model's strongest ability to recognize emotional cues. Figure 3 demonstrates the superior performance of the multimodal fusion strategy under the Perceiver IO architecture in the literature appreciation task, achieving both stable convergence and accurate identification of different types of literary elements.

E. TEACHING-ORIENTED VISUAL SEMANTIC INTERPRETATION MODULE

1) Fusion of Attention Weight Extraction and Projection Mechanism Design

To visualize the semantic attention paths within the model, this module uses the attention weight tensor generated by the Perceiver IO cross-attention mechanism, combined with the upstream and downstream semantic output structures, to perform layer-by-layer inversion and quantification of the attention levels of multimodal input tokens. Perceiver IO internally uses a query-key-value mapping mechanism to generate the attention matrix $A \in R^{Q\times K}$, where Q is the number of query vectors and K is the number of cross-modal key vectors. In the visualization module, the submatrix $A_{text} \in R^{Q\times N}$ for the text modality token in the attention matrix is extracted, where N represents the number of text tokens, as the basis for the model's attention to the local semantics of the text. Subsequently, the A_{text} output by each layer is average pooled according to the context window and normalized to generate a token-level interpretable score

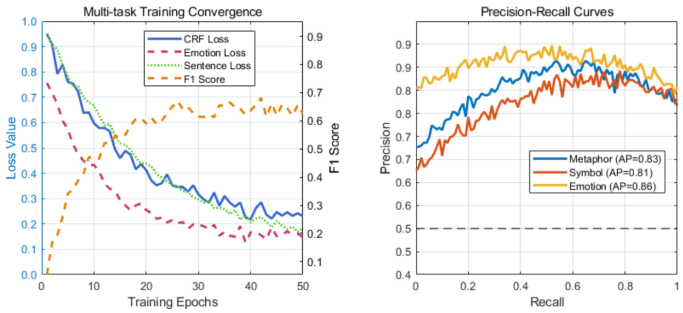

FIGURE 3. Multi-task Literary Feature Recognition Results (The figure caption suggests the following addition: The horizontal gray dashed line (Precision = 0.5) in the figure represents the performance baseline of the random classifier. All curves (metaphor, symbol, emotion) are significantly above this baseline, indicating that the model's recognition ability is significantly better than random guessing on all tasks.)

vector $\alpha = Softmax(meanpool(A_{text})) \in R^N$, which is used to describe the degree of focus of the model on each text segment in a specific input. To address the issue of assessing the importance of perceptual details, the WAAS indicator is introduced. This indicator combines the attention value of a token with the gradient contribution of the final sentiment classification output to the token, and is defined as follows:

$$WAAS_i = \alpha_i \cdot \left| \frac{\partial L_{emotion}}{\partial h_i} \right| \tag{10}$$

α_i represents the focus level of the ith token, and $\frac{\partial L_{emotion}}{\partial h_i}$ represents the gradient contribution of the corresponding token in the emotion subtask. This weight reflects the dual value of the token in perception and decision-making, making it suitable for constructing a visual interpretation matrix for teaching reference.

2) Visual Mapping and Teaching Assistance Interface Design

To ensure that the visualization results are intuitive and helpful in actual teaching scenarios, the system has designed a multi-level presentation structure. First, token-level WAAS values are linearly stretched and mapped to a color gradient space. The background of the token in the original text is colored to indicate the degree of attention the model pays to it. Areas of high attention are marked with high-saturation color blocks, and significant emotional turning points are indicated by color changes. In the interface presentation, sentence-level emotion category annotations are superimposed with key sentence identification labels to form a text attention word cloud. This word cloud allows teachers and learners to intuitively understand the model's focus path.

Figure 4 is a text attention heatmap, visually displaying the attention level (WAAS score) of different words in a literary text through a word cloud. Words such as "sank" (score 0.9), "ancient" (0.85), and "sun" (0.8) have high weights, indicating their key role in understanding textual semantics and expressing sentiment. Functional words such as "the" and "to" have lower weights, reflecting the model's focus on content words. This visualization helps reveal how the model identifies important information in the text, enabling it to more accurately capture literary elements and emotional characteristics, aligning with the paper's research theme of textual semantic interpretation and multi-task recognition.

This paper proposes a multimodal English literature listening comprehension model based on Perceiver IO. The core of the model consists of a five-stage architecture, as shown in Figure 5: First, Wav2Vec2.0 and BERT are used to extract audio frame features and text embeddings, respectively. The DTW algorithm and cross-modal attention are then used to achieve speech-text semantic alignment. Secondly, the standardized audio (256×768), text (128×768) and image features (64×768) are spliced into a unified representation vector (448×768); the core module uses Perceiver IO for four-layer iterative fusion, realizes multimodal interaction through cross-attention, and constructs context hierarchical modeling through self-attention; the output layer performs rhetorical element sequence labeling, sentiment three-classification and key sentence detection in parallel; finally, a multimodal heat map and WAAS teaching explanation interface are generated based on the attention weight, providing a visual decision-making basis for literary appreciation. The entire architecture achieves deep collaboration across the entire process, from heterogeneous data input to teaching applications.

III. FUSION PERFORMANCE EVALUATION IN ENGLISH LITERATURE LISTENING

The dataset used in this study is primarily designed for tasks in the multimodal context of English literature listening, encompassing three types of heterogeneous modal information: speech, text, and images. The speech data was selected from a subset of literature and excerpts of literary-themed speeches, ensuring a high degree of diversity in speaking speed, intonation, and emotional expression. The text is manually transcribed and annotated with literary elements by professional teachers, using the BIO encoding format to identify core information units such as metaphors, symbols, and emotions. The image modality mainly consists of video screenshots and handout illustrations to assist in understanding the text content. The entire dataset covers approximately 800 short literary listening passages with a total duration of approximately 42 hours. It contains 5,500 emotion-labeled

FIGURE 4. Text Attention Word Cloud

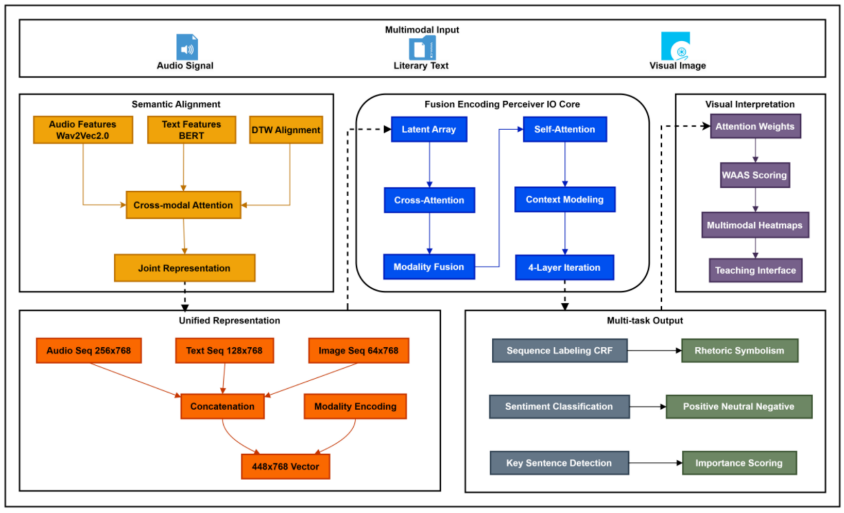

FIGURE 5. A Multimodal English Literature Listening Comprehension Model Based on Perceiver IO

samples and 4,700 rhetorical structure annotations. It is divided into training, validation, and test sets in a ratio of 7:2:1. This dataset provides real, complex, and cross-modal English literary context support for model evaluation and performance verification in this study.

A. MULTIMODAL SEMANTIC UNDERSTANDING AND RECOGNITION PERFORMANCE

To evaluate the model's performance in recognizing semantic elements in English literary listening materials under multimodal input conditions, accuracy and F1-score were selected as primary evaluation metrics. Using manually annotated English literary listening materials, the experiment compared the model's output on tasks such as sentiment classification and rhetoric recognition at the classification level, measuring the relationship between the number of correct classifications and the total number of predictions. The input modality is kept consistent during training to ensure that the impact of semantic fusion is not contaminated by other variables. Testing on multiple sample sets verifies the stability of the model's semantic fusion performance across varying literary styles, text complexity, and speech rate. Overall performance metrics measure the Perceiver IO architecture's ability to accurately perceive and recognize literary semantic information after fusing heterogeneous modal information. The selected models include Proposed Method, Transformer, LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network), MTL (Multi-

Task Learning), MMBT (Multimodal Bitransformer), ViLBERT (Vision-and-Language BERT), SpeechBERT (Speech Bidirectional Encoder Representations from Transformers), CLIP (Contrastive Language-Image Pre-training), ALBEF (Align Before Fuse). The baseline model is MMBT.

Figure 6 compares the performance of different models in multimodal semantic understanding tasks. The left figure shows that proposed model achieves the best performance in sentiment classification (accuracy/F1 score: 0.86/0.85), significantly outperforming the Transformer (0.72/0.71) and LSTM (0.68/0.66), demonstrating its superior multimodal fusion and sentiment discrimination capabilities. The right figure shows that proposed model achieves F1 scores above 0.80 in rhetoric recognition across five literary styles, demonstrating superior stability compared to the baseline model. In particular, the F1 score improves by 0.11 on dramatic text, demonstrating its ability to understand complex semantic structures and its good cross-genre generalization.

B. SPEECH-TEXT ALIGNMENT

This part uses the Dynamic Time Warping (DTW) algorithm to quantify the time step difference between speech frame sequences and text token sequences, measuring the model's ability to align pronunciation rhythm and intonation in English listening content. The DTW metric captures matching trends despite fluctuations in speaking rate. It also combines the BLEU (Bilingual Evaluation Understudy) score to assess the linguistic similarity between the speech transcription and the reference text. This dual-index system is used to determine whether the correspondence between speech input and text semantics is accurately mapped. In particular, when faced with unstable recitation rhythm and emotional expression in literary audio, it can significantly reflect the model's ability to align time-sensitive semantic information and evaluate its robustness in handling non-standard language and synchronous processing efficiency. Figure 7 compares the performance of different models on the speech-to-text alignment task. The left figure shows the distribution of the normalized DTW distance (a measure of temporal alignment accuracy) for each model. The proposed model achieves the lowest DTW median distance (approximately 0.2) and the most concentrated distribution, significantly outperforming the DTW-Align, CTC, and Seq2Seq models, demonstrating improved alignment accuracy and stability. The right figure shows the BLEU scores (a measure of text generation quality) at different speaking rates. Proposed model performs best at all speaking rates (normal: 0.82, fast: 0.78, variable: 0.80), significantly outperforming the comparison models (DTW-Align up to 0.75, CTC up to 0.72, Seq2Seq up to 0.68), demonstrating its enhanced robustness and semantic preservation capabilities in multi-rate environments.

C. MULTIMODAL COLLABORATIVE INFORMATION CONTRIBUTION

This section uses ablation experiments to evaluate the model's semantic understanding performance when different modal inputs are removed. The goal is to quantify the contribution of each modality to integrated contextual understanding. Based on the complete model (audio, text, and image), incomplete versions are constructed, removing any modality, and their performance is compared on the same task test set. This paper focuses on observing the changes in F1 scores across subtasks such as rhetoric recognition, symbol extraction, and sentiment assessment, and measuring fusion gain.

Figure 8 illustrates the contribution of multimodal collaboration to sentiment classification and rhetoric recognition tasks. The left panel of Figure 8 analyzes the impact of each modality on sentiment recognition by comparing the F1 scores of the "full model" with those of the "no audio," "no text," and "no vision" modal ablation conditions. Under the complete model, the F1 scores of positive, negative, and neutral emotions reached 0.86, 0.82, and 0.84, respectively, demonstrating a relatively high recognition effect. When the text modality was removed, the F1 scores of the three emotion categories dropped to 0.68, 0.65, and 0.72, respectively, with the most significant decrease, indicating that text information plays a dominant role in emotion classification. The right panel of Figure 8 shows the recognition capabilities of metaphor, symbolism, and personification under different modal conditions. The F1 scores for the full model are 0.85, 0.82, and 0.80, respectively. However, when audio or text is removed, performance degrades to varying degrees. The symbolism recognition score in the "no text" scenario is only 0.68, demonstrating the core value of linguistic modality in symbolism recognition.

In addition to modality ablation, we further conducted attribution analysis on typical errors in the rhetoric recognition task. Among 200 false negative samples, 68% stemmed from a lack of contextual support for cultural allusions (e.g., "Pandora's box"); 22% were due to label conflicts caused by nested rhetoric (e.g., "ironic metaphor"); and 10% were due to background music interfering with audio feature extraction. False positives mainly occurred when function words were mislabeled as core rhetoric elements (e.g., misclassifying "once" as "repetition"), accounting for approximately 15%. These analyses indicate that the model is robust to explicit rhetoric (e.g., personification, hyperbole), but there is still room for improvement in recognizing implicit cultural context and complex rhetoric, which points the way for future integration of external knowledge bases.

D. LITERARY ELEMENT RECOGNITION ACCURACY

Literary element recognition accuracy evaluates the model's ability to extract specific literary elements (such as symbolism, imagery, and metaphor) during the output phase. Relevant text segments are encoded as structured labels using the BIO sequence annotation format. During the evaluation process, the F1 score and recall rate for each literary element category are calculated against manual annotation standards to verify the model's robustness and discriminative ability in fine-grained semantic recognition tasks. This step focuses

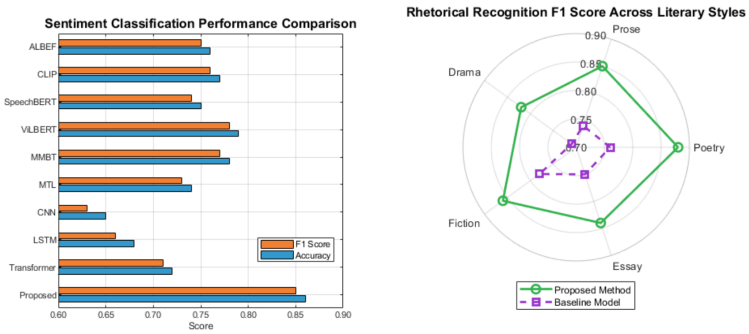

FIGURE 6. Performance Comparison of Different Models in Sentiment Classification and Rhetoric Recognition

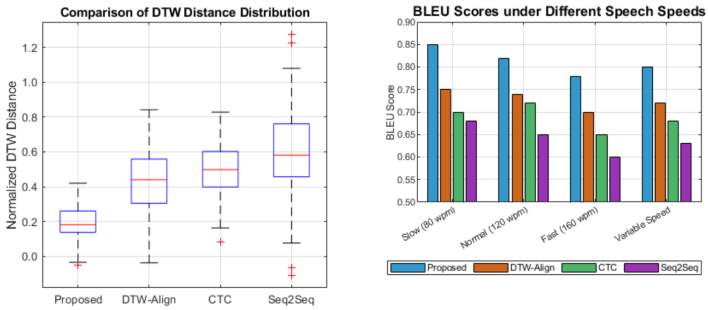

FIGURE 7. Speech-to-Text Alignment Performance Comparison: Cross-Model and Speech Rate Analysis

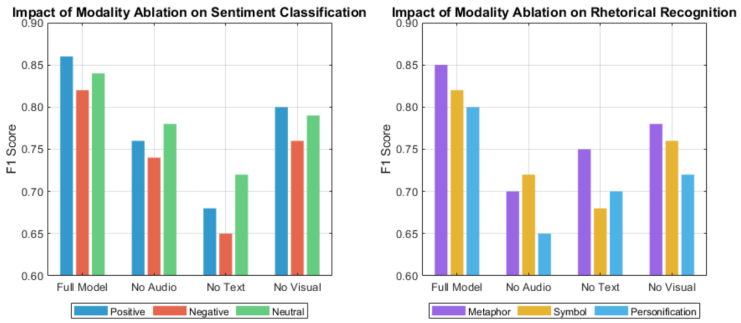

FIGURE 8. Contribution of Multimodal Collaboration to Sentiment Classification and Rhetoric Recognition Tasks

on testing the model's performance at the abstract semantic level, specifically whether it can improve the recognition accuracy of high-level semantic units with the support of multimodal fusion output. By analyzing the performance distribution across categories, it can identify which elements of appreciation are easily missed during contextual fusion, providing guidance for subsequent optimization.

Table 1 shows the model's performance in identifying various literary elements, focusing on its ability to understand fine-grained semantics. Overall, the model achieved high F1 scores of 0.85, 0.83, and 0.81 for identifying literary elements such as metaphor, personification, and repetition, respectively, demonstrating its accurate capture of these elements with typical semantic characteristics. Meanwhile, while the F1 scores for elements such as symbolism, imagery, and rhyme were slightly lower, they still remained between 0.77 and 0.82, demonstrating the model's strong robustness and generalization capabilities. The macro-average F1 value in Table 1 is 0.80, demonstrating the model's stable performance in the overall literary element recognition task.

TABLE 1. Literary Element Recognition Performance

Literary Element	Recall	F1 Score	Support Samples
Metaphor	0.84	0.85	325
Symbolism	0.81	0.82	287
Imagery	0.77	0.78	241
Allusion	0.78	0.80	196
Irony	0.74	0.75	173
Personification	0.81	0.83	264
Hyperbole	0.74	0.76	157
Alliteration	0.77	0.79	132
Repetition	0.78	0.81	205
Rhyme	0.75	0.77	178
Macro Average	0.78	0.80	2158

E. TEACHING INTERPRETABILITY SCORING

To evaluate the model's practical interpretability in teaching assistance applications, a subjective scoring system based on evaluations by teaching experts was designed. Five university English teachers conducted a blind review and rating of the interpretable outputs generated by the attention heatmap, sentiment visualization layer, and key sentence prompt module. The ratings covered focus accuracy (whether the attention placement was appropriate), explanation clarity (ease of understanding), and teaching applicability (whether it had pedagogical value). Scores for each dimension were normalized to a uniform scale, and the average was taken as the interpretability score. This evaluation aimed to measure the acceptability and potential of the model outputs in real-world teaching scenarios, verifying their effectiveness in assisting students in understanding complex literary contexts and improving teaching efficiency.

Table 2 shows the subjective ratings of the model's interpretability by teaching experts, covering focus accuracy, explanation clarity, and teaching applicability. Within focus accuracy, emotion focus scored the highest, with an average score of 4.76, indicating that the model excels in locating emotion-related information. Rhetorical focus scored slightly lower, at 4.62, suggesting room for improvement in refining the interpretation of linguistic rhetoric. In terms of clarity of explanation, the heatmap readability score was 4.60, indicating that the model-generated visualization aids were easy to understand. However, the clarity of the labels was slightly lower, averaging 4.40, suggesting that the accuracy of the labels still needs to be improved. Teaching applicability received the highest score, reaching 4.80 overall, with a particularly high score of 4.86 for classroom application, demonstrating that experts highly recognized the model's practical value and effectiveness in actual teaching scenarios. The overall composite score of 4.67, with a small standard deviation of 0.15, indicates relatively consistent expert evaluations, confirming the model's strong explanatory power and potential for practical application in assisting literature instruction.

TABLE 2. Expert Ratings on Teaching Interpretability

Evaluation Dimension	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Average Score	Standard Deviation
Focus Accuracy	4.7	4.6	4.8	4.9	4.5	4.7	0.15
- Emotional Focus	4.8	4.7	4.9	4.8	4.6	4.76	0.12
- Rhetorical Focus	4.6	4.5	4.7	4.9	4.4	4.62	0.19
Explanation Clarity	4.5	4.4	4.6	4.7	4.3	4.5	0.16
- Heatmap Readability	4.6	4.5	4.7	4.8	4.4	4.6	0.16
- Labeling Clarity	4.4	4.3	4.5	4.6	4.2	4.4	0.16
Teaching Applicability	4.8	4.7	4.9	5	4.6	4.8	0.17
- Classroom Application	4.9	4.8	5	4.9	4.7	4.86	0.11
- Student Feedback	4.7	4.6	4.8	4.9	4.5	4.7	0.16
Overall Score	4.67	4.57	4.77	4.83	4.47	4.67	0.15

To directly evaluate the model's performance on the core task of "literary appreciation," we invited five university teachers with doctoral degrees in English literature to conduct blind reviews of the appreciation reports generated by the model. The evaluation included: (1) the accuracy of the grasp of the emotional tone; (2) the reasonableness of the functional explanation of the rhetorical devices; and (3) the appropriateness of the interpretation of the cultural connotations of key images. Each expert scored the model based on a 5-point Likert scale. The results showed that the model's average score was 4.72 for "emotional tone," 4.58 for "rhetorical explanation," and 4.41 for "cultural imagery," with a total score of 4.57 (standard deviation 0.14). This indicates that although "literary appreciation" is subjective, the model's output has high credibility from an expert perspective, effectively supporting its application value in teaching scenarios.

IV. CONCLUSION

This study addresses the multimodal understanding challenges of specialized terminology and aesthetic appreciation in English listening comprehension tasks by building a unified speech-text-image fusion architecture based on Perceiver IO. By integrating Wav2Vec 2.0 and BERT to achieve semantic alignment, and combining multimodal unified representation with a cross-attention mechanism, the model effectively enhances its comprehensive understanding of literary rhetoric, emotional cues, and contextual structure. Supported by a multi-target prediction design and a visual interpretation module, the model demonstrates excellent practicality in teaching assistance and intelligent evaluation. Experiments demonstrate that proposed model performs exceptionally well in listening context modeling and literary aesthetic analysis, achieving a sentiment classification accuracy of 0.86 (F1=0.85), an F1 exceeding 0.80 for cross-genre rhetoric recognition, and a speech alignment distance approaching 0.2 (BLEU reaching a maximum of 0.85). Text modality contributes significantly, with a macro F1 of 0.80 for literary element recognition. However, the current model still relies heavily on the quality of training data and lacks robustness to diverse accents or complex background noise. Future work can combine large-scale pre-trained speech models with reinforcement learning mechanisms to further enhance the model's adaptability and generalization capabilities, thereby expanding its application in practical teaching

environments like specialized engineering and technical training.

AUTHOR CONTRIBUTIONS

Jing Li designed, collected and analyzed the data, and drafted the manuscript. Jing Li conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jing Li participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, Henan Provincial Higher Education Teaching Reform Research and Practice

Project Approval (Progressive Integration, Categorized Expansion: Research and Practice on the Paradigm of English Curriculum Serving the Career Development of Medical Students,2024SJGLX0876).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] P. Zhang and Y. Peng, "Multimodal reading in reading-only versus reading-while-listening modes: Evidence from Chinese language learners," Chinese as a Second Language Research, vol. 13, no. 2, pp. 215-236, 2024, doi: 10.1515/caslar-2024-2003.

[2] H. Cartner and D. Cameron, "Investigating metacognitive strategy awareness for multimodal listening," E-Learning and Digital Media, vol. 20, no. 5, pp. 424-441, 2023, doi: 10.1177/20427530221108014.

[3] M. Shamsitdinova, "Developing Multimodal Listening Skills in English for Specific (or special) Purposes: A pedagogical framework," The American Journal of Social Science and Education Innovations, vol. 6, no. 06, pp. 15-21, 2024, doi: 10.37547/tajssei/Volume06Issue06-03.

[4] S. Kaowiwattanakul, "Multimodal literature and CEFR reading proficiency: Improving literary reading skills in EFL learners," Indonesian Journal of Applied Linguistics, vol. 15, no. 1, pp. 20-34, 2025, doi: 10.17509/ijal.v15i1.75921.

[5] R. Masinde, D. Barasa, and L. Mandillah, "Effectiveness of using multimodal approaches in teaching and learning listening and speaking skills," Nairobi Journal of Humanities and Social Sciences, vol. 7, no. 2, pp. 91-112, 2023, doi: 10.58256/njhs.v7i2.1333.

[6] B. E. Smith, N. Amgott, and I. Malova, ""It made me think in a different way": Bilingual students' perspectives on multimodal composing in the English language arts classroom," Tesol Quarterly, vol. 56, no. 2, pp. 525-551, 2022, doi: 10.1002/tesq.3064.

[7] N. Trisanti, D. Suherdi, and D. Sukyadi, "Multimodality reflected in EFL teaching materials: Indonesian EFL in-service teacher's multimodality literacy perception," Language Circle: Journal of Language and Literature, vol. 17, no. 1, pp. 129-139, 2022, doi: 10.15294/lc.v17i1.38776.

[8] N. A. Jamil and A. A. Aziz, "The use of multimodal text in enhancing students' reading habit," Malaysian Journal of Social Sciences and Humanities (MJSSH), vol. 6, no. 9, pp. 487-492, 2021, doi: 10.47405/mjssh.v6i9.977.

[9] E. Tour and M. Barnes, "Engaging English language learners in digital multimodal composing: Pre-service teachers' perspectives and experiences," Language and Education, vol. 36, no. 3, pp. 243-258, 2022, doi: 10.1080/09500782.2021.1912083.

[10] K. V. Alvionita, L. Widyaningrum, and A. Prayogo, "EFL learners' reflection on digitally mediated multimodal project-based learning: Multimodal enactment in a listening-speaking class," Language Circle: Journal of Language and Literature, vol. 17, no. 1, pp. 87-97, 2022, doi: 10.15294/lc.v17i1.36466.

[11] F. V. Lim and L. Unsworth, "Multimodal composing in the English classroom: Recontextualising the curriculum to learning," English in Education, vol. 57, no. 2, pp. 102-119, 2023, doi: 10.1080/04250494.2023.2187696.

[12] P. Aedo and C. Millafilo, "Increasing vocabulary acquisition and retention in EFL young learners through the use of multimodal texts (memes)," Colombian Applied Linguistics Journal, vol. 24, no. 2, pp. 251-269, 2022, doi: 10.14483/22487085.18312.

[13] A. von Zansen, R. Hilden, and E. Laihanen, "The multimodal listening test in a high-stakes context: Gender-neutral or not?," International Journal of Listening, vol. 36, no. 2, pp. 152-170, 2022, doi: 10.1080/10904018.2021.1993446.

[14] E. Salamanti, D. Park, N. Ali, and S. Brown, "The efficacy of collaborative and multimodal learning strategies in enhancing English language proficiency among ESL/EFL Learners: A quantitative analysis," Research Studies in English Language Teaching and Learning, vol. 1, no. 2, pp. 78-89, 2023, doi: 10.62583/rseltl.v1i2.11.

[15] S. Hartle, R. Facchinetti, and V. Franceschi, "Teaching communication strategies for the workplace: A multimodal framework," Multimodal Communication, vol. 11, no. 1, pp. 5-15, 2022, doi: 10.1515/mc-2021-0005.

[16] C. Tyrer, "The voice, text, and the visual as semiotic companions: An analysis of the materiality and meaning potential of multimodal screen feedback," Education and Information Technologies, vol. 26, no. 4, pp. 4241-4260, 2021, doi: 10.1007/s10639-021-10455-w.

[17] S. Salsabila, N. Nurami, R. T. Oktaviana, S. Muljanto, and A. Kurnia, "Seeing How Multimodal Source Are Used in EFL Classroom," English Education and Applied Linguistics Journal (EEAL Journal), vol. 6, no. 2, pp. 102-106, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:269469218.

[18] L. Jiang, M. M. Gu, and F. Fang, "Multimodal or multilingual? Native English teachers' engagement with translanguaging in Hong Kong TESOL classrooms," Applied Linguistics Review, vol. 15, no. 4, pp. 1299-1319, 2024, doi: 10.1515/applirev-2022-0062.

[19] A. Corbitt, M. Wargo J, and C. O'Connor, "Encountering unnatural E-literature: tracing interpretation and relationality across multimodal response and digital annotation," English in Education, vol. 56, no. 2, pp. 186-200, 2022, doi: 10.1080/04250494.2021.1933424.

[20] C. Smith, "Deconstructing innercirclism: A critical exploration of multimodal discourse in an English as a foreign language textbook," Discourse: Studies in the Cultural Politics of Education, vol. 44, no. 1, pp. 88-105, 2023, doi: 10.1080/01596306.2021.1963212.

[21] K. A. Sherwani and M. K. Harchegani, "The Impact of Multimodal Discourse Analysis on the Improvement of Iraqi EFL Learners' Reading Comprehension Skill," Journal of Tikrit University for Humanities, vol. 29, no. 12, 2, pp. 1-19, 2022, doi: 10.25130/jtuh.29.12.2.2022.22.

[22] C. M. Ponzio and M. R. Deroo, "Harnessing multimodality in language teacher education: Expanding English-dominant teachers' translanguaging capacities through a multimodalities entextualization cycle," International Journal of Bilingual Education and Bilingualism, vol. 26, no. 8, pp. 975-991, 2023, doi: 10.1080/13670050.2021.1933893.

[23] A. N. Kluger, M. Lehmann, H. Aguinis, and G. Itzchakov, "A meta-analytic systematic review and theory of the effects of perceived listening on work outcomes," Journal of Business and Psychology, vol. 39, no. 2, pp. 295-344, 2024, doi: 10.1007/s10869-023-09897-5.

[24] C. Jia, K. F. Hew, and M. Li, "Towards a flipped SEF-ARCS decoding model to improve foreign language listening proficiency," Computer Assisted Language Learning, vol. 38, no. 3, pp. 369-396, 2025, doi: 10.1080/09588221.2023.2191655.

[25] Y. Zhao and V. Aryadoust, "An automatized semantic analysis of two large-scale listening tests: A corpus-based study," Language Testing, vol. 42, no. 3, pp. 312-343, 2025, doi: 10.1177/02655322241288598.

[26] M. Mpumuje, G. Bazimaziki, and J. D. L. P. Muragijimana, "Exploring the Role of Oral Literature in Enhancing Learners' Language Proficiency: A Case of Three Selected Secondary Schools in Rwanda," African Journal of Empirical Research, vol. 5, no. 2, pp. 752-763, 2024, doi: 10.51867/ajernet.5.2.65.