scientific reports

Check for updates

OPEN

Multimodal Alzheimer's disease recognition from image, text and audio

Byounghwa Lee[1]✉, Hwa Jeon Song[1], Young-Jin Park[2] & Byung Ok Kang[1]

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that significantly affects cognitive function. One widely used diagnostic approach involves analyzing patients' verbal descriptions of pictures. While prior studies have primarily focused on speech- and text-based models, the integration of visual context is still at an early stage. This study proposes a novel multimodal AD prediction model that integrates image, text, and audio modalities. The image and text modalities are processed using a vision-language model and structured as a bipartite graph before fusion, while all three modalities are integrated through a combination of co-attention-based intermediate fusion and late fusion, enabling effective inter-modality cooperation. The proposed model achieves an accuracy of 90.61%, outperforming state-of-the-art models. Furthermore, an ablation study quantifies the contribution of each modality using Shapley values, which serve as the foundation for a novel auxiliary loss function that adaptively adjusts modality importance during training. The findings indicate that integrating image, text, and audio modalities via a co-attention-based intermediate fusion enhances AD classification performance. Additionally, this study analyzes modality-specific attention patterns and key linguistic tokens, demonstrating that audio and text provide complementary cues for AD classification.

Keywords Alzheimer's disease, Artificial intelligence, Co-attention, Image-text-audio, Multimodal, Shapley value

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline, particularly memory impairment. Early detection is crucial for mitigating disease progression. Recent advancements in artificial intelligence (AI) have facilitated the development of automated approaches for AD detection, particularly leveraging speech and text analysis. One widely used cognitive assessment task is the Cookie Theft picture description task, a component of the Boston Diagnostic Aphasia Examination (BDAE)[1]. This task requires participants to spontaneously describe all elements in a picture, with responses recorded as audio files. A well-known dataset for this task is the Pitt corpus[2]. A subset of this corpus, the ADReSSo Challenge dataset[3], is age- and gender-matched and serves as a benchmark for AD detection. The ADReSSo dataset consists solely of audio recordings and is used to classify individuals into two groups: healthy control (HC) and AD.

Previous studies on AD recognition using picture description tasks have primarily focused on either audio-based or text-based models. Audio-based approaches extract acoustic features from speech signals using traditional signal processing techniques or deep learning embeddings[4,5]. Text-based methods, on the other hand, transcribe speech into text and leverage pre-trained language models for feature extraction[6,7]. More recent multimodal approaches integrate both audio and text through feature fusion or attention-based mechanisms, leading to improved classification performance[8,9]. Additionally, studies have incorporated external feedback from large language models (LLMs) such as ChatGPT and Mistral 7B, treating generated feedback as an auxiliary feature[10–12].

Despite these advancements, an essential component of the picture description task remains underutilized: the visual information itself. AD patients exhibit deficits in visual attention, object recognition, and scene interpretation, all of which are integral to describing an image. Incorporating image-based features can thus enhance the assessment of cognitive impairment. While prior work has made substantial progress using speech and text, how patients engage with visual stimuli—an essential aspect of the picture description task—has received relatively less attention. Given that visual deficits are often observed in early AD, incorporating image-based information may offer complementary insights for assessment.

[1]Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea. [2]Electro-Medicine Device Research Division, Korea Electrotechnology Research Institute, Ansan 15588, Republic of Korea. ✉email: byounghwa.lee@etri.re.kr

To address this limitation, we propose a novel multimodal AD prediction model that integrates image, text, and audio modalities. The proposed approach leverages a vision-language model (VLM) to construct a bipartite graph of image-text relationships, which is subsequently processed using a graph convolutional network (GCN) to learn structured representations. These graph embeddings are then fused with text and audio embeddings extracted from BERT and wav2vec2.0, respectively, facilitating effective inter-modality cooperation.

To facilitate fine-grained cross-modal interaction, we introduce a pairwise co-attention module that performs intermediate fusion by allowing each modality to attend to the others. This mechanism captures bidirectional dependencies and leads to improved alignment between modalities. Experimental results demonstrate that the co-attention design yields consistent performance gains across all evaluation metrics.

To quantify the contribution of each modality, we conduct an ablation study and compute Shapley values, which provide a robust measure of each modality's marginal contribution to model performance. Furthermore, we introduce a novel auxiliary loss function that adjusts modality-specific contributions based on the computed Shapley values. While its impact is modest compared to the co-attention module, it offers complementary gains and further refines the fusion process.

The proposed model achieves an accuracy of 90.61%, outperforming existing state-of-the-art (SOTA) models and demonstrating the benefits of incorporating visual information alongside traditional speech- and text-based AD recognition. Additionally, we analyze modality-specific attention patterns and identify key linguistic tokens, revealing that audio and text provide complementary cues for AD classification.

The main contributions of this paper are summarized as follows:

- A multimodal AD prediction model integrating image, text, and audio is proposed, achieving an accuracy of 90.61% with the co-attention module, and surpassing existing SOTA models. To the best of our knowledge, this study is the first to jointly incorporate all three modalities for AD recognition.
- We introduce a co-attention-based intermediate fusion module that enhances cross-modal alignment through pairwise modality interactions and improves overall classification accuracy.
- A bipartite graph is constructed to represent image-text relationships using a VLM and processed through a GCN to capture inter-modality dependencies. The learned graph representations are then fused with text and audio embeddings extracted from BERT and wav2vec2.0, enhancing multimodal integration.
- An ablation study is conducted to systematically evaluate all modality combinations, and Shapley values are computed to quantify each modalitymarginal contribution to classification performance.
- A novel Shapley-based auxiliary loss function is introduced to adjust modality-specific contributions during training, leading to improved prediction accuracy while preserving the distinct contributions of each modality.
- Modality-specific attention patterns are analyzed to interpret model decisions, demonstrating that audio and text provide complementary cues for AD classification.

Related work

Multimodal model for dementia detection

Multimodal dementia detection often integrates audio and text modalities from spontaneous speech. Audio features can be directly extracted from speech signals or represented as log-Mel spectrograms. Prior studies have explored various fusion strategies to enhance classification accuracy. For instance, one study[13] incorporated x-vectors, prosody, and emotion embeddings alongside word embeddings, achieving 80.30% accuracy. Another study[14] introduced WavBERT, where wav2vec outputs were transformed into BERT inputs to retain non-semantic information, improving accuracy to 83.10%.

Other studies have investigated different fusion strategies. A study[15] combined BERT with multiple acoustic models, such as x-vectors and encoder–decoder ASR embeddings, achieving 84.51% accuracy. Another approach[16] leveraged full transcripts as prompts for Whisper-based speech recognition, mitigating the limitations of short-segment training and reaching the same accuracy. Meanwhile, research[9] employed co-attention, deep context, and label smoothing, integrating BERT-encoded text with log-Mel spectrograms processed by a Data-efficient Image Transformer (DeiT), achieving 85.35% accuracy.

LLMs have also been explored for dementia detection. One study[10] incorporated ChatGPT-generated feedback as an auxiliary opinion feature, improving accuracy to 87.32%. Similarly, another experiment[11] tested Mistral 7B on the ADReSS dataset, reporting an accuracy of 81.3%.

While these studies focus primarily on audio and text modalities, recent work has highlighted the importance of visual context in cognitive assessment. Research[17] integrated images with descriptive text using pre-trained vision-language models such as CLIP (Contrastive Language-Image Pre-training)[18], demonstrating enhanced accuracy. The highest-performing model to date[19] constructed a bipartite graph using BLIP (Bootstrapping Language-Image Pre-training)[20] to capture image-text relationships and applied a graph neural network for classification, achieving 88.73% accuracy. Another approach introduced an explainable vision-language framework that transforms textual assessments into numeric scores to facilitate downstream classification[21]. These findings indicate that, in addition to the valuable linguistic and acoustic cues captured by audio and text modalities, incorporating visual information holds further potential for advancing dementia detection—particularly by enabling the analysis of deficits in visual attention and scene interpretation.

Multimodal fusion in deep learning

Multimodal fusion is categorized into early, intermediate, and late fusion[22–24]. Early fusion integrates raw input features into a single representation but requires precise alignment and results in high-dimensional inputs. Intermediate fusion combines modality-specific features at a shared layer, balancing interaction and independence but requiring careful design to prevent overfitting. Late fusion processes modalities independently

using pre-trained models, integrating features before prediction. While it limits low-level interactions, it allows flexible architectures. Since our model utilizes pre-trained feature extractors for each modality, we incorporate both intermediate and late fusion strategies—co-attention enables fine-grained cross-modal interaction, while late fusion ensures modular integration.

Shapley values for multimodal fusion

Shapley values[25] are widely utilized in machine learning to interpret feature contributions in predictive models. Originally introduced in game theory, the Shapley value quantifies contribution of each player in a coalition game. This concept remains valid when players are substituted with modalities. That is, when multiple modalities cooperate to form a model, the contribution of each modality to performance of the model can be assessed[26,27]. The Shapley value of modality m_i is defined as follows:

$$\phi_{m_i}(V) = \sum_{S \subseteq M \setminus \{m_i\}} \frac{|S|!(m - |S| - 1)!}{m!} \left(V(S \cup \{m_i\}) - V(S)\right), \tag{1}$$

where M denotes the set of modalities, m is the total number of modalities, V represents the performance metric (accuracy in this study), and S is a subset of M that excludes the modality m_i.

Vision-language models

Vision-language models (VLMs) are pre-trained models designed to jointly learn from image and text data. Notable examples include CLIP[18], BLIP[20], BLIP-2[28]. These models align image-text representations through contrastive learning and other optimization techniques, enabling efficient multimodal understanding.

CLIP employs contrastive learning to optimize the similarity between correctly paired image-text representations while minimizing mismatches. BLIP enhances image-text relationship modeling using objectives such as image-text contrastive loss (ITC), image-text matching loss (ITM), and language modeling (LM), along with CapFilt, a filtering method that improves dataset quality. BLIP-2 introduces Q-Former, a querying transformer that connects a frozen image encoder and a frozen LLM, optimizing both comprehension and generation capabilities while improving computational efficiency.

Proposed method

The proposed model integrates audio, text, and graph-based modalities to classify AD and HC participants based on the picture description task. Audio recordings are either used directly or transcribed into text, forming two separate input streams. Each modality is first processed independently, after which pairwise interactions are modeled through a co-attention-based intermediate fusion module. This module allows each modality to attend to the others, enhancing cross-modal alignment before final integration. The resulting representations are then combined using late fusion for classification. The overall framework is shown in Fig. 1, and the co-attention module is detailed separately in Fig. 2.

Representation extraction process for each modality

Audio modality

Audio recordings are sampled at 16,000 Hz, with a maximum duration of 20 seconds per segment for wav2vec2.0[29]. Longer recordings are split into overlapping segments, processed independently, and concatenated to form the final audio feature.

For instance, a 30-second sample is split into 20-second and 10-second segments, generating wav2vec2.0 outputs of dimensions $\mathbb{R}^{1000\times768}$ and $\mathbb{R}^{500\times768}$, respectively, which are concatenated to form a $\mathbb{R}^{1500\times768}$ feature. However, token length discrepancies across modalities pose alignment challenges, as the number of audio tokens is, on average, 25.8 times greater than text tokens in the ADReSSo dataset.

To mitigate this issue, audio tokens are averaged in 10-token chunks (200 ms), with a 100 ms overlap. This downsampling strategy balances resolution and efficiency while preserving speech dynamics relevant to AD classification[30,31]. The final audio feature for sample s is denoted as $x_{a,s} \in \mathbb{R}^{l_{a,s}\times768}$, where $l_{a,s}$ is the number of audio tokens. In Fig. 1, the process related to the audio modality is highlighted in yellow.

Text modality

The transcribed text is processed using Whisper[32] and encoded with BERT[33]. Due to BERT's 512-token limit, longer texts are segmented and processed independently, with the resulting features concatenated to form the final text representation.

BERT-encoded text embeddings capture semantic, contextual, and syntactic information. The final text feature for sample s is denoted as $x_{t,s} \in \mathbb{R}^{l_{t,s}\times768}$, where $l_{t,s}$ represents the number of text tokens. In Fig. 1, the process corresponding to the text modality is represented in light blue.

Graph modality

A bipartite graph is constructed to integrate image and text modalities using a VLM. A GCN[34] is then trained on this graph to extract multimodal representations. While the graph is constructed from image-text interactions, we treat it as a third modality for downstream classification and fusion.

The Cookie Theft image is first colorized and cropped into 10 sub-images based on Grad-CAM[35] analysis of the VLM[19]. Each sub-image is processed through a VLM to obtain image embeddings, while text is segmented at the sentence level and encoded using the same VLM.

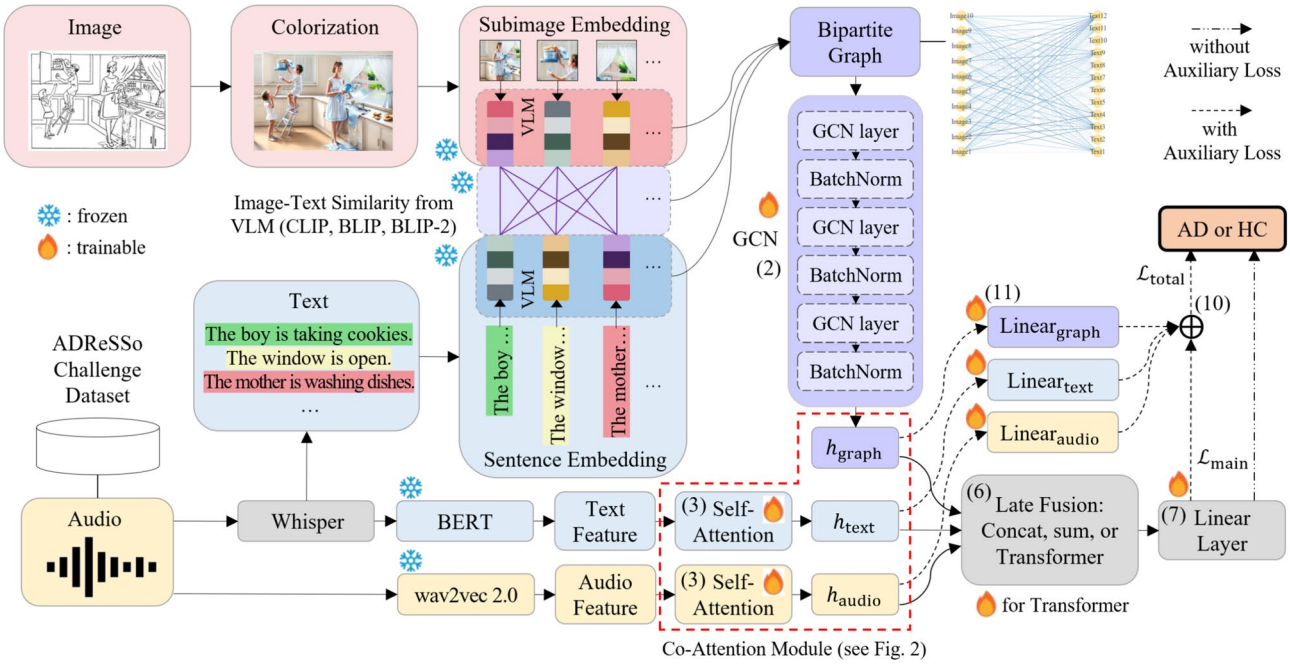

Fig. 1. Multimodal model for Alzheimer's disease recognition integrating image, text, and audio. Red boxes indicate image-related processes, blue represents text-related processes, and yellow corresponds to audio-related processes. Graph-based components integrating image and text are in purple. Parentheses indicate equation numbers. The red dashed box is replaced by a co-attention module designed for intermediate fusion.

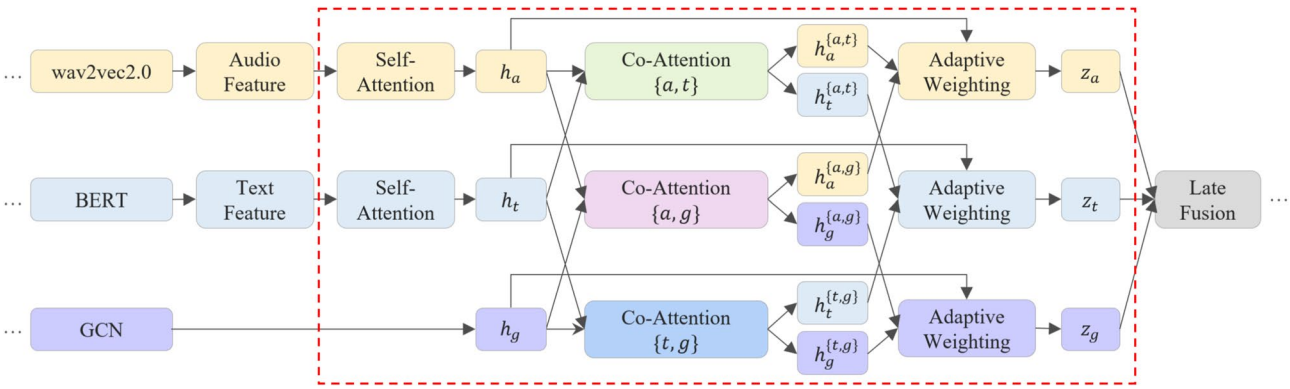

Fig. 2. A co-attention-based intermediate fusion module replaces the red dashed component in Fig. 1. Each modality, except for the graph, is first processed through a self-attention mechanism. Then, all three modalities (including the graph) undergo pairwise co-attention with one another. The resulting outputs are integrated via adaptive weighting to construct modality-specific representations (z_a, z_t, and z_g). These representations are subsequently processed by the late fusion module. If an auxiliary loss is applied, additional connections are directed to the three dedicated linear layers introduced in (11).

Cosine similarity between image and text embeddings determines edge weights in the bipartite graph. This enables the model to capture intricate relationships between visual and textual descriptions. For instance, a description may appear linguistically coherent but reference nonexistent entities in the image. Such semantic misalignments are detected by analyzing joint multimodal representations[19].

Graph nodes are assigned as either image-type or text-type nodes, maintaining a bipartite structure. The GCN processes this graph through three layers, with batch normalization applied after each layer. The propagation rule is defined as:

$$\mathrm{x}_{g,i}^{l} = \mathrm{W}_1^{l}\mathrm{x}_{g,i}^{l-1} + \mathrm{W}_2^{l}\sum_{j\in\mathcal{N}(i)} e_{j,i}\cdot\mathrm{x}_{g,j}^{l-1}, \tag{2}$$

where $x_{g,i}^{l}$ and $x_{g,i}^{l-1}$ represent the node embeddings for node *i* at layer *l* and $(l-1)$, respectively. If node *i* is an image node, then node *j* must be a text node, and vice versa. The edge weight from source node *j* to target node *i* is denoted as $e_{j,i}$, while $\mathcal{N}(i)$ represents the set of neighboring nodes of node *i*. The parameters W_1^l and W_2^l are learnable parameters.

The graph is processed through three GCN layers, with batch normalization applied after each layer. After passing through the first layer, the embedding dimension is reduced to a predefined size h_{dim}. With each successive layer, the hidden dimension is halved, resulting in a final vector of size $h_{dim}/4$. Consequently, the graph representation is represented as $h_g \in \mathbb{R}^{h_{dim}/4}$ for each sample. In Fig. 1, processes associated with the graph modality are marked in light purple.

Beyond three GCN layers, oversmoothing degrades performance, causing node embeddings to converge to similar representations. Empirical evaluation on the ADReSSo dataset indicates that a three-layer GCN achieves optimal performance[19].

The VLMs used are BLIP, BLIP-2, and CLIP. BLIP and BLIP-2 generate 768-dimensional embeddings with a maximum token length of 256, while CLIP produces 512-dimensional embeddings with a 77-token limit. For longer text inputs, sentences are segmented accordingly. CLIP also allows for image encoder selection, with ViT-B/16 chosen for its superior fine-grained feature extraction.

Intermediate fusion via self- and co-attention

For clarity, h_i denotes the self-attention feature of modality *i*, while z_i represents the feature after pairwise co-attention and adaptive weighting.

Self-attention encoding for each modality

To enrich modality-specific representations before applying cross-modal co-attention, self-attention is first applied to the audio feature ($x_{a,s}$) and text feature ($x_{t,s}$). This step enhances key structural and contextual signals within each modality. Following self-attention, global average pooling is performed along the sequence length dimension. A projection layer then reduces the feature dimension from 768 to $h_{dim}/4$ to align with the graph representation space. The resulting modality-specific representations for audio and text are defined as:

$$h_a = SA_a(x_{a,s}), \quad h_t = SA_t(x_{t,s}), \tag{3}$$

where each vector is $h_a \in \mathbb{R}^{h_{dim}/4}$ and $h_t \in \mathbb{R}^{h_{dim}/4}$. The self-attention modules for audio and text operate independently with non-shared parameters.

Pairwise co-attention across modalities

To enhance inter-modal alignment and capture richer interactions among modalities, we incorporate a co-attention-based intermediate fusion mechanism into the model architecture. Specifically, the fusion block highlighted by the red dashed box in Fig. 1 is replaced with the co-attention module illustrated in Fig. 2.

This module operates on the self-attention outputs of each modality. Audio and text features (h_a, h_t) are obtained from the self-attention encoders described earlier, while the graph representation h_g is derived via a GCN from the image-text bipartite graph.

We then apply pairwise co-attention across all modality pairs (audio-text, audio-graph, and text-graph) to explicitly model cross-modal dependencies. The general formulation is as follows:

$$\begin{aligned}
&\text{Co-Attention}(a,t) = h_a^{\{a,t\}} \text{ and } h_t^{\{a,t\}}, && \text{where } h_a^{\{a,t\}} = \text{MHA}(a,t,t) \text{ and } h_t^{\{a,t\}} = \text{MHA}(t,a,a)\\
&\text{Co-Attention}(a,g) = h_a^{\{a,g\}} \text{ and } h_g^{\{a,g\}}, && \text{where } h_a^{\{a,g\}} = \text{MHA}(a,g,g) \text{ and } h_g^{\{a,g\}} = \text{MHA}(g,a,a)\\
&\text{Co-Attention}(t,g) = h_t^{\{t,g\}} \text{ and } h_g^{\{t,g\}}, && \text{where } h_t^{\{t,g\}} = \text{MHA}(t,g,g) \text{ and } h_g^{\{t,g\}} = \text{MHA}(g,t,t).
\end{aligned} \tag{4}$$

Here, $\text{MHA}(q,k,v)$ denotes the multi-head attention operation, where the query (*q*), key (*k*), and value (*v*) inputs are each linearly projected before attention is computed. For example, $h_a^{\{a,t\}}$ represents the output of co-attention where audio attends to text (audio as query, text as key/value), while $h_t^{\{a,t\}}$ corresponds to text attending to audio.

The co-attention mechanism enables the model to capture fine-grained, bidirectional contextual dependencies between modalities. Unlike simple concatenation approaches, co-attention allows the model to learn task-specific cross-modal interactions dynamically during training.

Adaptive integration of attention representations

Following the co-attention layers, we apply adaptive weighting to integrate the original modality representation and the co-attended features. The modality-specific representations z_a, z_t, z_g are computed as follows:

$$\begin{aligned}
z_a &= w_1^a h_a + w_2^a h_a^{\{a,t\}} + w_3^a h_a^{\{a,g\}}\\
z_t &= w_1^t h_t + w_2^t h_t^{\{a,t\}} + w_3^t h_t^{\{t,g\}}\\
z_g &= w_1^g h_g + w_2^g h_g^{\{a,g\}} + w_3^g h_g^{\{t,g\}}
\end{aligned} \tag{5}$$

The weights $w_i^{m_i}$ for each modality $m_i \in \{a,t,g\}$ are computed by applying a softmax over learnable scalar scores associated with each component (original and co-attention representations), ensuring they sum to 1. Importantly, the original modality representation (e.g., h_a) serves as an identity path that helps retain modality-specific information. Since co-attention tends to emphasize relational signals between modalities, it may dilute

distinctive unimodal cues. By incorporating both the raw and co-attended features, the model learns to balance intra- and inter-modality information via training.

The resulting modality representations z_a, z_t, z_g are then passed into the late fusion module as inputs, as shown in (4) of Fig. 1. When the auxiliary loss is enabled, each z_{m_i} is also connected to its own classification head. Aside from the modified fusion block, the remaining components of the model remain identical to the original architecture.

Late fusion and AD classification process

Late fusion

To classify a given sample as AD or HC, modality-specific representations from audio, text, and graph are combined using a late fusion strategy. For the baseline model, these are denoted as h_a, h_t, and h_g (from self-attention or GCN), while for the co-attention model, the refined representations are z_a, z_t, and z_g from (5).

The late-fused representation h_f or z_f is defined as:

$$\begin{aligned} &h_f = [h_a, h_t, h_g] \quad \text{or} \quad z_f = [z_a, z_t, z_g] && \text{for concatenation,} \\ &h_f = h_a + h_t + h_g \quad \text{or} \quad z_f = z_a + z_t + z_g && \text{for summation,} \\ &h_f = \text{Transformer}([h_a, h_t, h_g]) \quad \text{or} \quad z_f = \text{Transformer}([z_a, z_t, z_g]) && \text{for Transformer.} \end{aligned} \tag{6}$$

Concatenation and summation directly combine the feature vectors without learnable parameters, while the Transformer fusion employs a trainable encoder to refine inter-modality interactions. The Transformer module uses six layers, a single attention head, and a dropout rate of 0.1, which performed best on the ADReSSo dataset.

AD classification

The final fused representation, denoted as h_f for the baseline model (with self-attention only) and z_f for the co-attention-enhanced model, is passed through a linear layer for binary classification:

$$y = \text{Linear}(h_f) \quad \text{or} \quad y = \text{Linear}(z_f), \tag{7}$$

where y represents the output logits for AD and HC classification. The model is trained with cross-entropy loss to learn the binary decision boundary.

Shapley value for the marginal contribution of each modality

The degree to which each modality contributes to improving model performance differs significantly. To quantify this, the Shapley value measures the performance gain of each modality by evaluating all possible modality combinations. Consider a case with three modalities, where the complete modality set M is defined as $M = \{m_1, m_2, m_3\}$, and the total number of modalities is $m = 3$. In the context of our study, m_1, m_2, and m_3 correspond to the audio, text, and graph modalities, respectively. Applying this to the Shapley value formulation in (1), the Shapley value for modality m_1 can be expressed as follows:

$$\phi_{m_1}(V) = \tfrac{1}{3}\left[V(m_1) - V(\emptyset)\right] + \tfrac{1}{6}\left[V(m_1, m_2) - V(m_2)\right] + \tfrac{1}{6}\left[V(m_1, m_3) - V(m_3)\right] + \tfrac{1}{3}\left[V(m_1, m_2, m_3) - V(m_2, m_3)\right]. \tag{8}$$

The Shapley values for modalities m_2 and m_3 can be derived in a similar manner. Here, $\emptyset$ represents the empty set, and $V(\emptyset)$ corresponds to random guessing, estimated at 50.7% given the test set class distribution. In (8), the accuracy values for single- and dual-modality cases are obtained from ablation experiments.

The normalized Shapley value for modality m_i is computed as

$$S_{m_i} = \frac{1}{Z}\phi_{m_i}(V), \quad \text{where } Z = \sum_{i=1}^{3} \phi_{m_i}(V). \tag{9}$$

These normalized Shapley values mitigate instability during training.

Modality-specific auxiliary loss based on shapley values

To integrate Shapley values into the training process, a modality-specific auxiliary loss is introduced. When predictions are made using a single modality, the corresponding loss is weighted by its normalized Shapley value and incorporated into the total loss:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{main}} + \lambda \sum_{i=1}^{3} S_{m_i} \mathscr{L}_{m_i}, \tag{10}$$

where λ controls the trade-off between the main loss $\mathscr{L}_{\text{main}}$ (cross-entropy loss with all modalities) and the auxiliary losses $\mathscr{L}_{m_i}$ for individual modalities.

Each modality-specific loss $\mathscr{L}_{m_i}$ is computed by feeding the unimodal feature representation (h_i or z_i) into a dedicated linear layer Linear_i for that modality. This ensures that predictions from each modality are learned independently:

$$y_i = \text{Linear}_i(h_i) \quad \text{or} \quad y_i = \text{Linear}_i(z_i), \quad i \in \{a, t, g\}. \tag{11}$$

Depending on the model variant, the auxiliary loss is computed using either the self-attended features h_i or the co-attended features z_i. The cross-entropy loss between y_i and the ground truth label is then computed to obtain $\mathscr{L}_{m_i}$. Empirical evaluation on the ADReSSo dataset shows that setting $\lambda = 0.1$ yields optimal performance, effectively balancing modality contributions.

Experiments

To assess the contribution of each architectural component, we compare three model variants: (1) a vanilla model, (2) a model with the Shapley-based auxiliary loss, and (3) a model with the co-attention module.

Dataset

The ADReSSo Challenge dataset, a benchmark resource for AD detection, originates from the Pitt corpus and is designed to ensure balanced age and gender distribution. It consists of Cookie Theft picture description task recordings with corresponding ground truth labels. Table 1 presents key dataset statistics, including the average number of sentences per sample and word count per sentence. On average, the AD group produces more sentences than the HC group, primarily due to frequent short responses such as "Okay" and "Uh-huh." However, the HC group exhibits longer and more structured sentences, suggesting greater linguistic complexity.

Implementation details

The implementation of BLIP and BLIP-2 is conducted using the LAVIS library[36], while OpenAI's CLIP library is used for CLIP. These libraries extract embeddings for sub-images and sentences, computing image-text similarity. The ViT-B/16 model is utilized for sub-image feature extraction for CLIP. The bipartite graph is constructed with PyTorch Geometric (PyG)[37], and the graph convolutional network (GCN) is implemented using PyG's GraphConv model[38].

Shapley-based auxiliary loss

To assess the importance of each modality, an ablation study is conducted by evaluating different subsets of the three modalities. The full model incorporates all three modalities: audio, text, and graph. Additionally, experiments are performed using two-modality combinations, specifically audio-text, audio-graph, and text-graph. Single-modality models are also performed individually for audio, text, and graph. For the graph-based modality, experiments are conducted separately using CLIP, BLIP, and BLIP-2 to compare their effectiveness. The architecture is adjusted depending on the number and type of included modalities, and each representation is projected through the corresponding Linear_i for auxiliary loss computation. Based on the ablation study, the normalized Shapley values for each VLM, each late fusion method, and all three modalities are computed, as shown in Table 3. These values are then used as coefficients for the modality-specific loss terms in the auxiliary loss function defined in (10).

Experimental settings

All experiments, including those with the co-attention module and Shapley-based auxiliary loss, are conducted using PyTorch[39]. Training is performed on an NVIDIA RTX A6000 GPU, with an average runtime of 20 minutes per fold excluding feature extraction. The model is trained with a learning rate of $1e^{-6}$, a batch size of 4, a dropout rate of 0.2, and an auxiliary loss coefficient $\lambda = 0.1$. These hyperparameters were optimized through prior experiments. Training is conducted with a maximum of 2000 epochs, but early stopping is applied if validation loss does not improve for 250 consecutive epochs, based on preliminary experiments indicating that shorter patience values led to premature convergence. A 5-fold cross-validation (CV) strategy is employed. During the test phase, predictions from the five CV models are aggregated using a majority voting strategy to generate the final classification result.

Results

AD classification performance of the three modalities

Table 2 presents AD classification results for all modality configurations, including three-modality, two-modality, and single-modality settings. Additionally, the table reports performance across three VLMs (CLIP, BLIP, and BLIP-2) and three late fusion techniques: concatenation, summation, and Transformer fusion. Accuracy (Ac) is the primary evaluation metric, while precision (Pr), recall (Rc), F1-score (F1), and specificity (Sp) serve as secondary metrics.

Performance differences by modality configuration

The highest accuracy (90.14%) is obtained when using text and graph modalities without audio, but this is only observed with BLIP. For CLIP and BLIP-2, the best performance is achieved with all three modalities. This

		# Samples	# Sentences	# Avg. Words per Sentence
Train	HC	79	14.67 ± 7.55	13.48 ± 13.55
	AD	87	20.07 ± 16.18	10.40 ± 12.77
Test	HC	36	13.44 ± 7.06	12.89 ± 10.68
	AD	35	17.71 ± 8.96	8.22 ± 4.60

Table 1. ADReSSo dataset description.

Modalities	VLM	Late Fusion	Pr (%)	Rc (%)	F1 (%)	Sp (%)	Ac (%)
Audio, Text, and Graph	CLIP	Concat	90.32 ± 0.00	80.00 ± 0.00	84.85 ± 0.00	91.67 ± 0.00	85.92 ± 0.00
		Sum	92.40 ± 1.80	80.95 ± 1.65	86.29 ± 1.52	93.52 ± 1.60	87.32 ± 1.41
		Transformer	92.61 ± 1.25	82.86 ± 4.95	87.39 ± 2.14	93.52 ± 1.60	88.26 ± 1.63
	BLIP	Concat	93.46 ± 2.90	80.00 ± 2.86	86.15 ± 0.43	94.44 ± 2.78	87.32 ± 0.00
		Sum	93.41 ± 0.12	80.95 ± 1.65	86.73 ± 1.00	94.44 ± 0.00	87.79 ± 0.81
		Transformer	90.81 ± 0.16	84.76 ± 1.65	87.68 ± 0.96	91.67 ± 0.00	88.26 ± 0.81
	BLIP-2	Concat	92.57 ± 1.69	82.86 ± 0.00	87.44 ± 0.76	93.52 ± 1.60	88.26 ± 0.81
		Sum	91.69 ± 1.61	83.81 ± 1.65	87.56 ± 0.88	92.59 ± 1.60	88.26 ± 0.81
		Transformer	93.67 ± 2.95	83.81 ± 3.30	88.43 ± 2.38	94.44 ± 2.78	89.20 ± 2.15
	Average		92.33 ± 1.84	82.22 ± 2.67	86.95 ± 1.51	93.31 ± 1.77	**87.85 ± 1.30**
Text and Graph	CLIP	Concat	89.64 ± 3.39	81.90 ± 1.65	85.58 ± 2.16	90.74 ± 3.21	86.38 ± 2.15
		Sum	90.79 ± 4.40	81.90 ± 3.30	86.01 ± 0.25	91.67 ± 4.81	86.85 ± 0.81
		Transformer	90.12 ± 5.86	84.76 ± 1.65	87.31 ± 3.13	90.74 ± 5.78	87.79 ± 3.25
	BLIP	Concat	**94.68 ± 1.84**	84.76 ± 4.36	**89.41 ± 2.83**	**95.37 ± 1.60**	**90.14 ± 2.44**
		Sum	89.10 ± 1.81	**85.71 ± 2.86**	87.37 ± 2.31	89.81 ± 1.60	87.79 ± 2.15
		Transformer	90.51 ± 2.73	80.95 ± 1.65	85.43 ± 0.66	91.67 ± 2.78	86.38 ± 0.81
	BLIP-2	Concat	90.90 ± 2.66	84.76 ± 1.65	87.69 ± 0.66	91.67 ± 2.78	88.26 ± 0.81
		Sum	92.43 ± 3.68	80.95 ± 3.30	86.29 ± 3.09	93.52 ± 3.21	87.32 ± 2.82
		Transformer	91.78 ± 4.61	83.81 ± 4.36	87.56 ± 3.46	92.59 ± 4.24	88.26 ± 3.25
	Average		91.11 ± 3.45	**83.28 ± 3.04**	**86.96 ± 2.31**	91.98 ± 3.39	87.69 ± 2.20
Audio and Graph	CLIP	Concat	83.32 ± 5.25	74.29 ± 2.86	78.41 ± 0.89	85.19 ± 5.78	79.81 ± 1.63
		Sum	85.22 ± 2.52	76.19 ± 4.36	80.37 ± 1.97	87.04 ± 3.21	81.69 ± 1.41
		Transformer	83.33 ± 6.51	76.19 ± 5.95	79.60 ± 6.21	85.19 ± 5.78	80.75 ± 5.86
	BLIP	Concat	84.57 ± 2.64	78.10 ± 3.30	81.18 ± 2.40	86.11 ± 2.78	82.16 ± 2.15
		Sum	80.00 ± 0.00	80.00 ± 0.00	80.00 ± 0.00	80.56 ± 0.00	80.28 ± 0.00
		Transformer	86.32 ± 1.70	78.10 ± 1.65	82.00 ± 1.50	87.96 ± 1.60	83.10 ± 1.41
	BLIP-2	Concat	84.91 ± 4.80	80.00 ± 4.95	82.35 ± 4.37	86.11 ± 4.81	83.10 ± 4.23
		Sum	82.65 ± 5.09	79.05 ± 5.95	80.56 ± 0.54	83.33 ± 7.35	81.22 ± 0.81
		Transformer	84.89 ± 7.17	80.95 ± 5.95	82.53 ± 0.33	85.19 ± 8.49	83.10 ± 1.41
	Average		83.91 ± 4.19	78.10 ± 4.19	80.78 ± 2.68	85.19 ± 4.75	81.69 ± 2.56
Audio and Text	CLIP	Concat	93.47 ± 3.12	80.95 ± 1.65	86.74 ± 1.65	94.44 ± 2.78	87.79 ± 1.63
		Sum	91.52 ± 4.53	80.00 ± 2.86	85.29 ± 1.54	92.59 ± 4.24	86.38 ± 1.63
		Transformer	92.43 ± 1.56	80.95 ± 1.65	86.29 ± 0.24	93.52 ± 1.60	87.32 ± 0.00
	BLIP	Concat	93.41 ± 3.34	80.95 ± 3.30	86.73 ± 3.23	94.44 ± 2.78	87.79 ± 2.93
		Sum	91.52 ± 4.53	80.00 ± 2.86	85.29 ± 1.54	92.59 ± 4.24	86.38 ± 1.63
		Transformer	93.33 ± 0.00	80.00 ± 0.00	86.15 ± 0.00	94.44 ± 0.00	87.32 ± 0.00
	BLIP-2	Concat	91.43 ± 1.66	80.95 ± 1.65	85.86 ± 0.90	92.59 ± 1.60	86.85 ± 0.81
		Sum	93.17 ± 0.27	78.10 ± 3.30	84.95 ± 2.09	94.44 ± 0.00	86.38 ± 1.63
		Transformer	93.33 ± 0.00	80.00 ± 0.00	86.15 ± 0.00	94.44 ± 0.00	87.32 ± 0.00
	Average		**92.62 ± 2.43**	80.21 ± 2.09	85.94 ± 1.47	**93.72 ± 2.26**	87.06 ± 1.36
Audio	-	-	74.58 ± 1.77	75.24 ± 1.65	74.88 ± 0.55	75.00 ± 2.78	75.12 ± 0.81
Text	-	-	89.84 ± 1.11	83.81 ± 4.36	86.66 ± 1.86	90.74 ± 1.60	87.32 ± 1.41
Graph	CLIP	-	85.24 ± 1.88	81.90 ± 4.36	83.46 ± 1.44	86.11 ± 2.78	84.04 ± 0.81
	BLIP	-	90.93 ± 2.34	**85.71 ± 2.33**	88.23 ± 2.09	91.67 ± 2.27	88.73 ± 1.99
	BLIP-2	-	87.56 ± 2.74	80.00 ± 0.00	83.59 ± 1.25	88.89 ± 2.78	84.51 ± 1.41

Table 2. Model performance across three-, two-, and single-modality configurations. Results include experiments with three VLMs (CLIP, BLIP, BLIP-2) and three late fusion methods (concatenation, summation, Transformer). Accuracy (Ac) is the primary metric, while precision (Pr), recall (Rc), F1 score (F1), and specificity (Sp) serve as secondary metrics. Each value represents the mean of three runs, with standard deviations reported. The highest value for each metric is in bold and underlined, while the second-highest is underlined. The highest average performance is in bold.

suggests that, in certain cases, audio fusion introduces noise, which negatively affects classification. The results of the Shapley value analysis in the following section confirm that audio contributes the least. However, excluding audio does not result in statistically significant improvements (p-value is 0.7101 for BLIP), and on average, it decreases accuracy by 0.17 pp. While the two-modality setting yields the highest accuracy, the three-modality setting provides more stable performance overall.

In contrast, excluding text results in a statistically significant accuracy drop of 6.16 pp (p-value is $1.223e^{-13}$), supporting the Shapley analysis that text is the most critical modality. Removing the graph modality reduces accuracy by 0.78 pp (p-value is 0.0391), indicating that image-text relationships enhance classification.

Performance differences by VLM and late fusion method

Among the three VLMs, BLIP-2 achieves the highest average accuracy, followed by BLIP and CLIP. BLIP-2 outperforms BLIP by 0.78 pp, and BLIP surpasses CLIP by 0.63 pp, though these differences are not statistically significant.

For late fusion, Transformer fusion yields the highest accuracy, followed by summation and concatenation. Transformer fusion outperforms summation fusion by 0.78 pp, and summation fusion surpasses concatenation by 0.63 pp. However, only the difference between Transformer and concatenation fusion is statistically significant (p-value is 0.0487).

Performance of single-modality models

Among single-modality models, text and graph-based models perform comparably, whereas the audio-only model exhibits the lowest performance. The graph model achieves the highest recall when using BLIP, but its accuracy remains lower than that of the text model. This indicates that while image-text relationships contribute to classification, semantic text features remain the most informative.

Shapley value results

The accuracy values computed in Table 2 are used as the function V in (8), and the corresponding Shapley values (ϕ) and normalized Shapley values (S) are reported in Table 3. These values are computed separately for each VLM and late fusion method. Across all configurations, the text modality consistently has the highest contribution, followed by the graph modality, with the audio modality contributing the least. This indicates that semantic information plays a crucial role in AD classification, while the relationship between image and text provides additional discriminative power. The audio modality, having already been partially converted into text through transcription, contributes less information. Furthermore, the lower performance of the audio modality compared to text has been consistently observed in previous AD recognition research. However, incorporating handcrafted audio features could potentially enhance its impact on classification performance.

The effect of the shapley-based auxiliary loss

The results of training with the Shapley-based auxiliary loss, constructed based on the normalized Shapley values from Table 3, are presented in Table 4. For comparison, the table also includes the results of the default model without auxiliary loss. The highest accuracy (89.20%) is obtained using Shapley-based auxiliary loss with BLIP-2 and concatenation fusion. This configuration also exhibits lower variance compared to Transformer fusion without auxiliary loss.

On average, applying auxiliary loss improves accuracy by 0.21 pp. Performance gains vary across VLMs, with CLIP and BLIP-2 improving by 0.94 pp and 0.32 pp, respectively. However, BLIP's accuracy decreases by 0.63 pp, indicating a suboptimal combination with Shapley-based auxiliary loss. For late fusion, concatenation and summation fusion improve by 0.78 pp and 0.32 pp, respectively, while Transformer fusion decreases by 0.47 pp. The decline in Transformer fusion performance suggests that the trainable parameters in the fusion module may not align well with Shapley-based auxiliary loss. Overall, the Shapley-based auxiliary loss enhances classification performance with minimal computational cost, requiring only three additional linear layers. Given its benefits, the proposed auxiliary loss model is a promising approach for multimodal AD recognition.

Late Fusion	Normalization	CLIP			BLIP			BLIP-2		
		(ϕ_a, S_a)	(ϕ_t, S_t)	(ϕ_g, S_g)	(ϕ_a, S_a)	(ϕ_t, S_t)	(ϕ_g, S_g)	(ϕ_a, S_a)	(ϕ_t, S_t)	(ϕ_g, S_g)
Concat	×	24.09	33.48	27.85	22.92	33.01	30.9	24.56	33.24	29.96
	O	0.277	0.386	0.321	0.264	0.38	0.356	0.283	0.383	0.345
Sum	×	24.48	33.17	29.17	23.31	33.17	30.82	24.48	33.64	29.64
	O	0.282	0.382	0.336	0.268	0.382	0.355	0.282	0.387	0.341
Transformer	×	24.48	34.1	29.18	24.56	32.31	30.9	24.95	33.64	30.11
	O	0.282	0.393	0.336	0.283	0.372	0.356	0.287	0.387	0.347

Table 3. The Shapley values (as defined in (8)) and normalized Shapley values (as defined in (9)) for each modality are defined as follows. Shapley values (ϕ_a, ϕ_t, ϕ_g) and normalized Shapley values (S_a, S_t, S_g) for audio, text, and graph modalities. Across all VLMs and fusion methods, text consistently has the highest Shapley value, while audio has the lowest.

Modalities	Auxiliary Loss	VLM	Late Fusion	Pr (%)	Rc (%)	F1 (%)	Sp (%)	Ac (%)
Audio, Text, and Graph	×	CLIP	Concat	90.32 ± 0.00	80.00 ± 0.00	84.85 ± 0.00	91.67 ± 0.00	85.92 ± 0.00
			Sum	92.40 ± 1.80	80.95 ± 1.65	86.29 ± 1.52	93.52 ± 1.60	87.32 ± 1.41
			Transformer	92.61 ± 1.25	82.86 ± 4.95	87.39 ± 2.14	93.52 ± 1.60	88.26 ± 1.63
		BLIP	Concat	93.46 ± 2.90	80.00 ± 2.86	86.15 ± 0.43	94.44 ± 2.78	87.32 ± 0.00
			Sum	93.41 ± 0.12	80.95 ± 1.65	86.73 ± 1.00	94.44 ± 0.00	87.79 ± 0.81
			Transformer	90.81 ± 0.16	84.76 ± 1.65	87.68 ± 0.96	91.67 ± 0.00	88.26 ± 0.81
		BLIP-2	Concat	92.57 ± 1.69	82.86 ± 0.00	87.44 ± 0.76	93.52 ± 1.60	88.26 ± 0.81
			Sum	91.69 ± 1.61	83.81 ± 1.65	87.56 ± 0.88	92.59 ± 1.60	88.26 ± 0.81
			Transformer	93.67 ± 2.95	83.81 ± 3.30	**88.43 ± 2.38**	94.44 ± 2.78	**89.20 ± 2.15**
	O	CLIP	Concat	91.60 ± 1.69	82.86 ± 0.00	87.00 ± 0.76	92.59 ± 1.60	87.79 ± 0.81
			Sum	91.69 ± 1.61	83.81 ± 1.65	87.56 ± 0.88	92.59 ± 1.60	88.26 ± 0.81
			Transformer	93.48 ± 0.12	81.90 ± 1.65	87.30 ± 1.00	94.44 ± 0.00	88.26 ± 0.81
		BLIP	Concat	91.43 ± 1.66	80.95 ± 1.65	85.86 ± 0.90	92.59 ± 1.60	86.85 ± 0.81
			Sum	94.37 ± 1.79	79.05 ± 1.65	86.01 ± 0.25	**95.37 ± 1.60**	87.32 ± 0.00
			Transformer	92.37 ± 2.05	80.95 ± 3.30	86.28 ± 2.77	93.52 ± 1.60	87.32 ± 2.44
		BLIP-2	Concat	**94.59 ± 1.80**	82.86 ± 0.00	88.33 ± 0.78	**95.37 ± 1.60**	**89.20 ± 0.81**
			Sum	92.67 ± 1.52	83.81 ± 1.65	88.00 ± 0.21	93.52 ± 1.60	88.73 ± 0.00
			Transformer	90.29 ± 3.64	86.67 ± 4.36	88.34 ± 1.57	90.74 ± 4.24	88.73 ± 1.41

Table 4. Multimodal model with and without Shapley-based auxiliary loss. The graph modality represents data where image and text are tightly linked via a graph structure. For the proposed model, results are averaged over three runs, with standard deviations reported. Accuracy (Ac) is the primary metric, while precision (Pr), recall (Rc), F1 score (F1), and specificity (Sp) are secondary metrics. The highest value for each metric is in bold and underlined, and the second-highest is underlined.

Performance of the co-attention module compared to baseline models

The results of the co-attention module are presented in Table 5. For comparison, the table also reports the performance of representative multimodal baseline models. The proposed co-attention model, configured with BLIP-2 and either Transformer-based or summation-based late fusion, achieves an accuracy of **90.61%**. It outperforms all baseline models across key evaluation metrics, with a precision of 94.76%, recall of 87.62%, F1 score of 90.19%, and specificity of 95.37%, demonstrating consistent improvements over previous approaches.

We also conducted experiments with and without the auxiliary loss, and the model demonstrated consistently high performance in both cases. On average, the application of auxiliary loss led to a 0.06 pp improvement in accuracy; however, the difference was not statistically significant. Although the auxiliary loss is effective in the baseline model, its impact diminishes under co-attention, suggesting that cross-modal alignment is already well achieved through pairwise attention mechanisms.

Discussion

Explainability

Analysis of attention weights in each modality

Shapley values were used to assess each modality's contribution to classification performance. A natural follow-up question is which specific components within each modality have the most significant impact on performance. Although a precise analysis of this is challenging, visualizing the self-attention weights of both audio and text can provide insights into which audio and text tokens are most influential for classification.

Fig. 3 visualizes attention weight distributions for audio and text, where (a) corresponds to an HC sample and (b) to an AD sample. The top row in each figure represents the audio attention weights from passing the audio feature through the self-attention mechanism, followed by the corresponding Mel spectrogram and transcribed speech. The bottom row represents the text attention weights obtained when passing the text feature through self-attention.

Distinct attention patterns were observed for audio and text. In Fig. 3 (b), an AD sample, audio attention is primarily focused on "the boys are trying to get a cookie jar out," while text attention highlights "drying" in "ladies were drying dishes." This suggests that while text attention captures semantic content, audio attention encodes prosodic and acoustic features, such as intonation, pitch, and speech emphasis. The Pearson correlation coefficient between energy in the spectrogram and audio attention weights is 0.0595, indicating a negligible relationship. This confirms that audio attention weights capture information beyond raw energy levels, likely focusing on prosodic aspects instead.

For the graph modality, key components were identified by analyzing GCN-learned output vectors. Specifically, crucial keywords and sentences were extracted based on their similarity or dissimilarity to representative vectors of each group in a previous study[19].

Multimodal Baseline Model						
Modalities	**Architecture**	**Pr (%)**	**Rc (%)**	**F1 (%)**	**Sp (%)**	**Ac (%)**
Audio and Text	BERT and acoustic models[15]	92.00	74.00	83.00	94.00	84.51
Audio and Text	Co-attention with label smoothing[9]	84.43 ± 1.59	86.29 ± 4.19	85.27 ± 1.78	84.43 ± 2.19	85.35 ± 1.44
Audio and Text	Text, audio, and ChatGPT[10]	88.06	87.32	87.25	94.44	87.32
Text and Graph	GNN-VLM[19]	90.93 ± 2.34	85.71 ± 2.33	88.23 ± 2.09	91.67 ± 2.27	88.73 ± 1.99

Proposed Model - Co-Attention								
Modalities	**Auxiliary Loss**	**VLM**	**Late Fusion**	**Pr (%)**	**Rc (%)**	**F1 (%)**	**Sp (%)**	**Ac (%)**
Audio, Text, and Graph	×	CLIP	Concat	91.50 ± 1.78	81.90 ± 1.65	86.43 ± 1.52	92.59 ± 1.60	87.32 ± 1.41
			Sum	89.63 ± 1.21	81.90 ± 3.30	85.55 ± 1.22	90.74 ± 1.60	86.38 ± 0.81
			Transformer	93.54 ± 0.21	82.86 ± 2.86	87.86 ± 1.70	94.44 ± 0.00	88.73 ± 1.41
		BLIP	Concat	91.53 ± 1.56	81.90 ± 1.65	86.43 ± 0.24	92.59 ± 1.60	87.32 ± 0.00
			Sum	93.50 ± 2.97	80.95 ± 1.65	86.74 ± 0.69	94.44 ± 2.78	87.79 ± 0.81
			Transformer	91.62 ± 1.49	82.86 ± 2.86	86.99 ± 1.10	92.59 ± 1.60	87.79 ± 0.81
		BLIP-2	Concat	93.75 ± 0.00	85.71 ± 0.00	89.55 ± 0.00	94.44 ± 0.00	90.14 ± 0.00
			Sum	92.96 ± 1.54	**87.62 ± 1.65**	**90.19 ± 0.86**	93.52 ± 1.60	**90.61 ± 0.81**
			Transformer	**94.76 ± 1.75**	85.71 ± 0.00	90.00 ± 0.78	**95.37 ± 1.60**	**90.61 ± 0.81**
	○	CLIP	Concat	88.89 ± 2.75	82.86 ± 2.86	85.71 ± 0.86	89.81 ± 3.21	86.38 ± 0.81
			Sum	91.53 ± 1.56	81.90 ± 1.65	86.43 ± 0.24	92.59 ± 1.60	87.32 ± 0.00
			Transformer	92.80 ± 1.64	85.71 ± 0.00	89.11 ± 0.76	93.52 ± 1.60	89.67 ± 0.81
		BLIP	Concat	89.63 ± 1.21	81.90 ± 3.30	85.55 ± 1.22	90.74 ± 1.60	86.38 ± 0.81
			Sum	93.54 ± 3.03	81.90 ± 1.65	87.32 ± 1.67	94.44 ± 2.78	88.26 ± 1.63
			Transformer	90.81 ± 0.16	84.76 ± 1.65	87.68 ± 0.96	91.67 ± 0.00	88.26 ± 0.81
		BLIP-2	Concat	93.75 ± 0.00	85.71 ± 0.00	89.55 ± 0.00	94.44 ± 0.00	90.14 ± 0.00
			Sum	92.89 ± 1.49	86.67 ± 1.65	89.65 ± 0.17	93.52 ± 1.60	90.14 ± 0.00
			Transformer	**94.76 ± 1.75**	85.71 ± 0.00	90.00 ± 0.78	**95.37 ± 1.60**	**90.61 ± 0.81**

Table 5. Comparison of multimodal baseline models and the proposed co-attention model. The graph modality represents data where image and text are tightly linked via a graph structure. For the proposed model, results are averaged over three runs, with standard deviations reported. Accuracy (Ac) is the primary metric, while precision (Pr), recall (Rc), F1 score (F1), and specificity (Sp) are secondary metrics. The highest value for each metric is in bold and underlined, and the second-highest is underlined. The proposed co-attention model outperforms all baselines.

Analysis of key text tokens

To examine the role of specific text tokens, the mean text attention weight across all tokens was computed. Fig. 4 presents the top 30 tokens with the highest attention weights in HC and AD groups. In both HC and AD groups, tokens such as "##flow" (from "overflow"), "reaching", and "sink" consistently receive high attention weights. However, notable differences are observed between the groups. In the HC group, "mouth" and "wiping" receive high attention weights. The token "mouth" appears in descriptions of the girl placing her fingers near her mouth, and "wiping" is used to describe the mother wiping dishes with a cloth rather than washing them, aligning with the actual scene. (Nevertheless, "washing" appears as a high-attention token in both groups.)

Conversely, in the AD group, "picture" and "cooking" receive high attention. The token "picture", which frequently appears in the investigator's speech, indicates that the relative importance of AD participants' spoken words is lower compared to that of the investigator. Meanwhile, "cooking" is highlighted despite the absence of a cooking scene, reflecting the tendency of AD participants to inaccurately describe the picture.

These findings are consistent with prior research[40], which identified "overflow", "drying", "open", and "stool" as key lexical features distinguishing Primary Progressive Aphasia (PPA) patients from healthy controls. In our study, "open" appears exclusively in the HC group, whereas "##flow" and "drying" are identified as important tokens in both groups, aligning with prior research.

Limitations

This study introduces an intermediate fusion strategy based on pairwise co-attention to enable cross-modal interactions among audio, text, and image modalities. However, more sophisticated fusion mechanisms were not explored. In addition, the current framework does not incorporate handcrafted acoustic features (e.g., prosody, pitch, pauses), which may enhance the utility of the audio modality. Furthermore, the experiments are conducted solely on the ADReSSo dataset due to its alignment with the picture description task, but generalization to other datasets remains to be validated in future work.

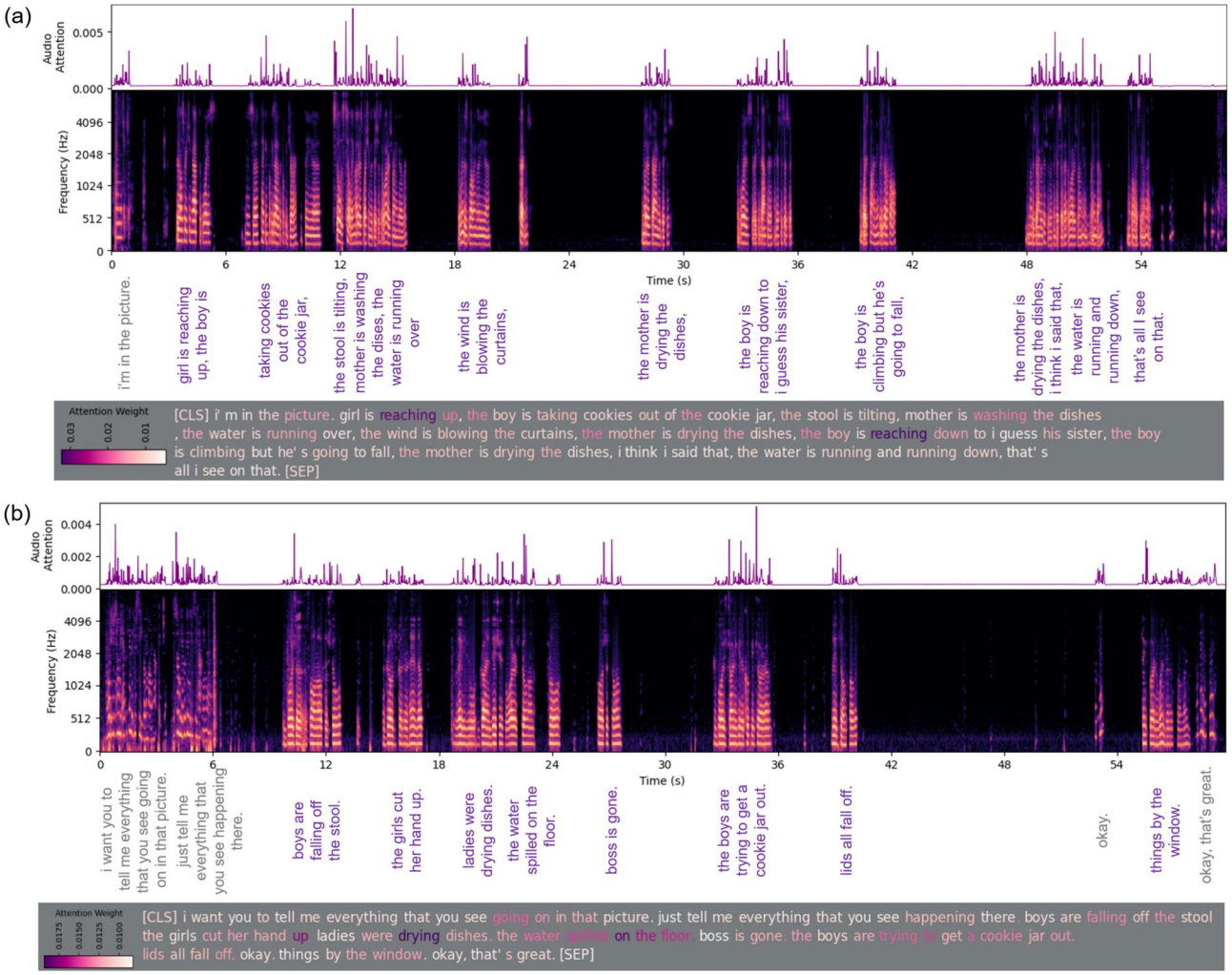

Fig. 3. Comparison of attention weights in audio and text. The top row visualizes self-attention weights from the audio encoder, followed by the Mel spectrogram and aligned transcription. Investigator speech is highlighted in gray, and participant speech in purple. The bottom row presents text self-attention weights. Darker purple indicates higher weights. (**a**) HC sample, (**b**) AD sample.

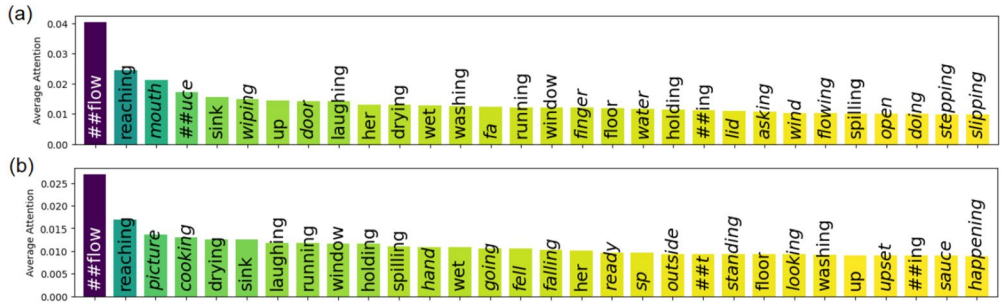

Fig. 4. Top 30 tokens with the highest average text attention weights for (**a**) HC and (**b**) AD groups. Tokens in italics appear exclusively in one group, while non-italicized tokens are common to both. The prefix "##" denotes a subword unit. In both groups, tokens such as "##flow" (from "overflow"), "reaching", and "sink" appear with high attention weights. However, notable differences are observed between the groups. HC samples emphasize "mouth" (describing the girl touching her mouth) and "wiping" (referring to dish-wiping instead of washing). AD samples highlight "picture" (frequent in investigator speech) and "cooking" (despite no cooking scene).

Conclusion

This study proposes a multimodal AD recognition framework that integrates image, text, and audio modalities. To enhance cross-modal alignment, a pairwise co-attention module is introduced, enabling intermediate fusion between modality pairs. This module significantly improves classification performance, achieving an accuracy of 90.61% and surpassing existing SOTA models. The framework also leverages vision-language models (VLMs) and a GCN to extract graph-based image-text representations, which are fused with BERT- and wav2vec2.0-based features. Ablation and Shapley value analyses quantify each modality's contribution. Based on this, a Shapley-based auxiliary loss is incorporated to enhance supervision, especially in the default model; its effect is less pronounced with co-attention, which already facilitates strong cross-modal integration. Finally, attention pattern analysis reveals that audio and text provide complementary cues, supporting the effectiveness of fine-grained multimodal integration for AD classification.

Future work could extend beyond binary classification by incorporating large language models (LLMs) into multimodal AD recognition frameworks, as recent surveys have highlighted their potential in medical imaging[41]. In addition, recent work has shown that combining fine-tuned LLMs with acoustic features can enhance AD detection performance[12], pointing to potential directions for advancing speech-language modeling for AD diagnosis. The target scope should also be broadened to include mild cognitive impairment (MCI) and related neurodegenerative disorders for broader clinical applicability.

Data availability

The data that support the findings of this study were obtained under license from DementiaBank (https://dementia.talkbank.org) and are not publicly available. Access to the data is restricted and managed by the administrators of DementiaBank. Interested researchers should consult the data access guidelines provided on the official DementiaBank website and contact the administrators as instructed. The authors do not have the authority to share the dataset directly.

Received: 14 April 2025; Accepted: 5 August 2025
Published online: 08 August 2025

References

1. Goodglass, H., Kaplan, E. & Weintraub, S. *BDAE: The Boston diagnostic aphasia examination* (Lippincott Williams & Wilkins Philadelphia, 2001).
2. Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J. & McGonigle, K. L. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. neurology* **51**, 585–594 (1994).
3. Luz, S., Haider, F., de la Fuente, S., Fromm, D. & MacWhinney, B. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint* arXiv:2104.09356 (2021).
4. Balagopalan, A. & Novikova, J. Comparing acoustic-based approaches for alzheimer's disease detection. *arXiv preprint* arXiv:2106.01555 (2021).
5. Gauder, M. L., Pepino, L. D., Ferrer, L. & Riera, P. Alzheimer Disease Recognition Using Speech-Based Embeddings From Pre-Trained Models. In *Proc. Interspeech 2021*, 3795–3799, :10.21437/Interspeech.2021-753 (2021).
6. Pan, Y. *et al.* Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech. In *Interspeech*, 3810–3814 (2021).
7. Syed, Z. S., Syed, M. S. S., Lech, M. & Pirogova, E. Tackling the adresso challenge 2021: The muet-rmit system for alzheimer's dementia recognition from spontaneous speech. In *Interspeech*, 3815–3819 (2021).
8. Touvron, H. *et al.* Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357 (PMLR, 2021).
9. Ilias, L. & Askounis, D. Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. *Knowledge-Based Syst.* **277**, (2023).
10. Bang, J.-U., Han, S.-H. & Kang, B.-O. Alzheimer's disease recognition from spontaneous speech using large language models. *ETRI Journal* (2024).
11. Botelho, C. *et al.* Macro-descriptors for alzheimer's disease detection using large language models. In *Interspeech 2024*, 1975–1979, :10.21437/Interspeech.2024-1255 (2024).
12. Casu, F., Lagorio, A., Ruiu, P., Trunfio, G. A. & Grosso, E. Integrating fine-tuned llm with acoustic features for enhanced detection of alzheimer's disease. *IEEE J. Biomed. Heal. Informatics* (2025).
13. Wang, N., Cao, Y., Hao, S., Shao, Z. & Subbalakshmi, K. Modular multi-modal attention network for alzheimer's disease detection using patient audio and language data. In *Interspeech*, 3835–3839 (2021).
14. Zhu, Y., Obyat, A., Liang, X., Batsis, J. A. & Roth, R. M. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In *Interspeech*, vol. 2021, 3790 (NIH Public Access, 2021).
15. Pappagari, R. et al. Automatic detection and assessment of alzheimer disease using speech and language technologies in low-resource scenarios. *In Interspeech* **2021**, 3825–3829 (2021).
16. Li, J. & Zhang, W.-Q. Whisper-based transfer learning for alzheimer disease classification: Leveraging speech segments with full transcripts as prompts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11211–11215 (IEEE, 2024).
17. Zhu, Y. *et al.* Evaluating picture description speech for dementia detection using image-text alignment. *arXiv preprint* arXiv:2308.07933 (2023).
18. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
19. Lee, B., Bang, J.-U., Song, H. J. & Kang, B. O. Alzheimer's disease recognition using graph neural network by leveraging image-text similarity from vision language model. *Sci. Reports* **15**, 997 (2025).
20. Li, J., Li, D., Xiong, C. & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900 (PMLR, 2022).
21. Casu, F., Grosso, E., Lagorio, A., Ruiu, P. & Trunfio, G. A. Leveraging multimodal vision language models for early detection of alzheimer's disease. In *2025 33rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP)*, 291–298 (IEEE, 2025).
22. Ramachandram, D. & Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* **34**, 96–108 (2017).

23. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **41**, 423–443 (2018).
24. Yin, S. *et al.* A survey on multimodal large language models. *arXiv preprint* arXiv:2306.13549 (2023).
25. Shapley, L. S. A value for n-person games. *Contribution to the Theory of Games* **2** (1953).
26. Hu, P., Li, X. & Zhou, Y. Shape: An unified approach to evaluate the contribution and cooperation of individual modalities. *arXiv preprint* arXiv:2205.00302 (2022).
27. Jeon, J., Kim, J., Park, J. & Kim, J. Msv: Contribution of modalities based on the shapley value. In *2024 IEEE International Conference on Consumer Electronics (ICCE)*, 1–6 (IEEE, 2024).
28. Li, J., Li, D., Savarese, S. & Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742 (PMLR, 2023).
29. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* **33**, 12449–12460 (2020).
30. Wu, S.-L., Kingsbury, E., Morgan, N. & Greenberg, S. Incorporating information from syllable-length time scales into automatic speech recognition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 2, 721–724 (IEEE, 1998).
31. Wu, S.-L., Kingsbury, B., Morgan, N. & Greenberg, S. Performance improvements through combining phone-and syllable-scale information in automatic speech recognition. *In ICSLP* **1**, 160–163 (1998).
32. Radford, A. *et al.* Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518 (PMLR, 2023).
33. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805 (2018).
34. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint* arXiv:1609.02907 (2016).
35. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).
36. Li, D. *et al.* Lavis: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 31–41 (2023).
37. Fey, M. & Lenssen, J. E. Fast graph representation learning with pytorch geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).
38. Morris, C. et al. Weisfeiler and leman go neural: Higher-order graph neural networks. *In Proceedings of the AAAI conference on artificial intelligence* **33**, 4602–4609 (2019).
39. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).
40. Henderson, S. K. *et al.* Lexical markers of disordered speech in primary progressive aphasia and 'parkinson-plus' disorders. *Brain Commun.* **6**, fcae433 (2024).
41. Wang, P. et al. Large language model for medical images: A survey of taxonomy, systematic review, and future trends. *Big Data Min. Anal.* **8**, 496–517 (2025).

Acknowledgements

This research was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No.CAP21054-300).

Author contributions

All authors conceived the experiment, B.L. conducted the experiment, all authors analysed the results, B.L. wrote the manuscript. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.