Unveil Multi-Picture Descriptions for Multilingual Mild Cognitive Impairment Detection via Contrastive Learning

Kristin Qi[1,*], Jiali Cheng[2,*], Youxiang Zhu[1], Hadi Amiri[2], Xiaohui Liang[1]
[1]Computer Science, University of Massachusetts, Boston, MA, USA
[2]Computer Science, University of Massachusetts, Lowell, MA, USA
Email: {yanankristin.qi001, Youxiang.Zhu001, xiaohui.liang}@umb.edu, {jiali_cheng, hadi_amiri}@uml.edu

Abstract—Detecting Mild Cognitive Impairment from picture descriptions is critical yet challenging, especially in multilingual and multiple picture settings. Prior work has primarily focused on English speakers describing a single picture (e.g., the 'Cookie Theft'). The TAUKDIAL-2024 challenge expands this scope by introducing multilingual speakers and multiple pictures, which presents new challenges in analyzing picture-dependent content. To address these challenges, we propose a framework with three components: (1) enhancing discriminative representation learning via supervised contrastive learning, (2) involving image modality rather than relying solely on speech and text modalities, and (3) applying a Product of Experts (PoE) strategy to mitigate spurious correlations and overfitting. Our framework improves MCI detection performance, achieving a +7.1% increase in Unweighted Average Recall (UAR) (from 68.1% to 75.2%) and a +2.9% increase in F1 score (from 80.6% to 83.5%) compared to the text unimodal baseline. Notably, the contrastive learning component yields greater gains for the text modality compared to speech. These results highlight our framework's effectiveness in multilingual and multi-picture MCI detection.

Index Terms—Mild cognitive impairment, multilingual and multimodal analysis, speech and language processing

I. INTRODUCTION

Mild Cognitive Impairment (MCI) is an intermediate stage between normal cognition and dementia. Early detection of MCI is crucial, as timely interventions before the onset of dementia symptoms can potentially delay disease progression. With the rapid growth of AI applications on the Internet of Things (IoT), these technologies make it possible to detect and monitor cognitive health biomarkers [1]–[3].

Traditional cognitive assessments rely on blood tests, neuroimaging, and in-clinic tests like the Mini-Mental State Examination (MMSE). While effective, these methods are time-consuming and labor-intensive. In contrast, AI-driven speech processing has emerged as a tool for early MCI detection through non-invasive and cost-effective screening. These processing systems can automatically identify biomarkers related to cognitive decline, such as hesitation gaps, reduced fluency, and changes in word richness [4].

The majority of previous research in dementia detection through speech focused on English-speaking cohorts who describe a single "Cookie Theft" picture in the Pitt Corpus [5]. Recently, the TAUKDIAL-2024 challenge includes English and Chinese to detect MCI through multilingual speech processing: each participant describes three different pictures in their respective languages [6]. However, this multilingual and multi-picture context presents new challenges that reduce the effectiveness of existing models developed for the "Cookie Theft" picture description task.

Specifically, TAUKDIAL-2024 introduces two major challenges. First, while studies processing linguistic features in "Cookie Theft" speech achieve 85.4% accuracy [7], performance on TAUKDIAL-2024 drops to 59.2% Unweighted Average Recall (UAR) [6], [8], [9]. This noticeable drop is due to increased variability between different pictures, which reduces accuracy using linguistic features. Previous approaches do not address complementary information in different pictures and their impact on the detection of cognitive decline. The second challenge involves the dataset's limited size and imbalanced class labels. Models developed on this speech data risk overfitting and relying on spurious feature correlations rather than true causal ones, which results in shortcut solutions with undermined reliability [10]–[12].

In this paper, we propose a framework with three components to address the above challenges. For the first challenge, we enhance the model's performance by incorporating picture variability through supervised contrastive learning [13]. This approach enhances picture-specific feature discrimination by pulling together descriptions of the same picture while pushing apart descriptions of different pictures. We also integrate image embeddings as an additional modality input. Both approaches leverage to improve model abilities with the complementary picture information. For the second challenge, we adopt a Product of Experts (PoE) strategy for modality concatenation to let multimodalities collectively determine the final prediction label. Since reliance on a single modality leads to performance degradation, PoE prioritizes causal features, and this multimodal decision-making approach mitigates spurious correlations and overfitting [10], [12].

In summary, our contributions are threefold:
- Our proposed framework leverages contrastive learning to enhance picture-specific discriminative features from multi-picture descriptions. We incorporate image embeddings as an additional modality to improve performance.
- We adopt Product of Expert (PoE) for multimodal processing to mitigate spurious correlations and reduce

subgroup disparities, which is crucial for models to generalize using the small-sized and class-imbalanced multilingual speech data.

- We validate our framework through extensive experiments. Experiments demonstrate a +7.1% improvement in UAR (68.1% to 75.2%) and a +2.9% improvement in F1 score (80.6% to 83.3%) compared to the text unimodal baseline. Notably, our contrastive learning approach enhances performance using text modality compared to speech modality. These results show that our framework effectively addresses the challenges of multilingual, multi-picture MCI detection.

II. RELATED WORK

A. Multilingual Studies in Detecting Cognitive Decline

Speech and language processing for cognitive decline detection spans several decades, with most methods focusing on English speakers. Two widely used speech data are the ADReSS challenge and Pitt Corpus [5]. Speech processing combines acoustic and linguistic features with machine learning classifiers. Studies have reported accuracy ranging from 72.9% to 89.6% depending on models [14], [15]. A few studies include more languages. The ADReSS-M challenge includes English and Greek speakers for dementia detection, achieving accuracy up to 87% [16]. Pérez-Toro *et al.* have included English, German, and Spanish speakers for Alzheimer's detection [17] and achieved UAR of 73%. Furthermore, the TAUKDIAL-2024 challenge focuses on MCI detection, and the established baseline achieves 59.18% UAR. Other reported UARs are between 49.4% and 56.18% [8], [9]. These results imply the challenges in developing highly performing and generalizable models for multilingual MCI detection.

B. Overfitting and Spurious Correlation

Recent work by Liu *et al.* [11] raised important questions about the reliability of reported high accuracy in dementia detection. Their analysis of the Pitt Corpus revealed that models could achieve good results by relying solely on silence segments, which is the Clever Hans effect. This finding suggests that models might be exploiting dataset-specific shortcuts, leading to spurious correlations rather than learning true causal patterns. Spurious correlations can come from imbalanced and noisy datasets. The consequence of spurious correlations is bias against minorities and over-reliance on data shortcuts, limiting real-world application deployment. While rectifying the weakness of relying on spurious correlation is crucial for clinical applications, it remains a challenge in adapting existing algorithms in medical routines. To address the challenges, we adopt PoE to mitigate reliance on spurious features and reduce performance disparities between subgroups.

III. DATASET

The TAUKADIAL-2024 challenge consists of speech recordings from the participants aged 60 to 90 years old (mean age: 72), with different numbers of male and female speakers. It includes English and Chinese-speaking participants and maintains three recordings for each participant, where each participant describes three different pictures.

The last digit in each recording file name indicates the specific picture being described, with 1, 2, and 3 corresponding to the first, second, and third pictures, respectively. Six pictures were described, three in English and three in Chinese. We relabeled each picture as 1-6 in our data processing.

For each participant, a standard clinical assessment using the MMSE was performed. Based on their MMSE scores, participants were labeled as either normal cognitive decline (NC) or MCI. The labels are provided in this dataset. The dataset comprises 129 participants (74 MCI and 55 NC), resulting in a total of 387 samples (222 MCI and 165 NC). Of these samples, 186 are English speakers and 201 are Chinese speakers. Details of the dataset are described in Table I.

We are aware that previous work [6], [8], [9], [18] used the same speech data and reported higher metrics under cross-validation settings; however, their performance on the test set remains substantially lower (below 60%). Since the data does not release the test set publicly, we report cross-validation performance only and do not compare against previous work on the test set.

TABLE I
DATASET STATISTICS WITH ENGLISH AND CHINESE SPEAKERS.

	Age	Male	Female	MMSE	English	Chinese
MCI	73.36	87	134	25.84	123	98
NC	71.85	63	102	29.07	63	102

IV. PROBLEM FORMULATION

Our framework integrates embeddings from speech, text, and picture inputs to classify participants as NC or MCI in multilingual settings. Suppose the dataset is $D = \{(X_i, Y_i)\}_{i=1}^{N}$ containing N speech recording samples. Each sample $X_i = (S_i, T_i, I_i)$ consists of three input modalities: speech (S_i), text (T_i), and image (I_i). The corresponding label $Y_i \in \{0, 1\}$ indicates NC or MCI, respectively. The task is to predict $\hat{Y}_i \in \{0, 1\}$ for each sample X_i.

V. METHODS

Our framework is shown in Figure 1. Embeddings from multimodality are combined into a joint representation. This representation is processed through the PoE and trained with supervised contrastive loss and cross-entropy loss to produce final labels. In this section, we explain the details of feature extraction, contrastive learning, and the PoE approach.

A. Preprocessing

For preprocessing speech inputs, we first removed the noisy background and applied a 16 kHz sampling rate to the audio with a single (mono) channel. Speech was transcribed using the Whisper large-v3[1] automatic speech recognition (ASR) model to convert audio recordings into text. We maintained the default feature dimension settings for pre-trained models.

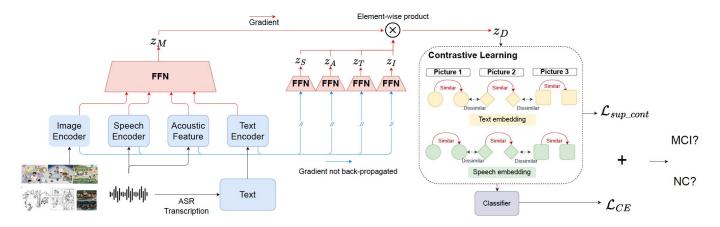[1] https://huggingface.co/openai/whisper-large-v3

Fig. 1. Details of components in our framework: (i) the left side shows feature extraction from each unimodality: speech, acoustic, text, and image, and they are processed through feed-forward networks (FFN). (ii) middle shows that PoE fuses these modalities through element-wise product or element-wise addition of their logarithms in log space to produce logits z_D. (iii) right side shows supervised contrastive learning to enhance picture-specific feature discrimination.

B. Feature Extraction

We use established models for extracting features from speech and transcriptions.

Speech Encoder: We process each speech input S_i through a speech encoder f_S to obtain speech embeddings $f_S(S_i)$. Specifically, we use the pre-trained Whisper-tiny model[2] [19], which extracts features from a transformer-based architecture.

Acoustic Features: The handcrafted acoustic features supplement previous speech encoder features. We extract a comprehensive set of acoustic features $f_A(S_i)$ using the DisVoice[3] package [20]. These handcrafted features are phonation characteristics, phonological information, articulation dynamics, and prosodic elements, which are relevant to identifying the MCI speech patterns.

Text Encoder: For each transcription T_i, linguistic embeddings $f_T(T_i)$ are obtained using the pre-trained language model. We use the BERT-base[4] [21] as the text encoder.

Image Encoder: For each audio recording where a participant describes a picture, we obtain image embeddings from the corresponding picture using ResNet-34 [22] as the image encoder f_I. Specifically, the encoder maps each picture being described I_i to its embedding representation $f_I(I_i)$.

C. Contrastive Learning for Feature Enhancement

To enhance discrimination between descriptions of different pictures, we employ contrastive learning. This approach learns to pull together positive pairs (similar descriptions) while pushing apart negative pairs (dissimilar descriptions). Ultimately, it aims to minimize distances between positive pairs while maximizing distances between negative pairs.

We adopt supervised contrastive learning [13] where pair relationships are determined by picture labels (see Section III). We define positive pairs as descriptions of the same picture and negative pairs as descriptions of different pictures.

[2] https://huggingface.co/openai/whisper-tiny
[3] https://github.com/jcvasquezc/DisVoice
[4] https://huggingface.co/google-bert/bert-base-uncased

Since positive pairs have similar embeddings and negative pairs have distant embeddings, the model learns to minimize distances between positive pair embeddings while maximizing distances between negative pair embeddings. Thus, contrastive learning enforces the learning of both picture-specific features and cognitive status features (NC vs. MCI). The equation of contrastive learning loss ($\mathcal{L}_{\text{sup\_cont}}$) is shown below:

$$\mathcal{L}_{\text{sup\_cont}} = \sum_{k \in \mathcal{B}} \frac{-1}{|P(k)|} \sum_{p \in P(k)} \log \frac{\exp\left(\frac{\mathbf{h}_k \cdot \mathbf{h}_p}{\tau}\right)}{\sum_{n \in N(k)} \exp\left(\frac{\mathbf{h}_k \cdot \mathbf{h}_n}{\tau}\right)},$$

(1)

where $\mathbf{h}_k$ denotes the embedding of a description in the batch. $\mathcal{B}$ represents the set of batches. $P(k)$ is the set of descriptions of the same picture (positive) as indices k. $N(k)$ is the set of descriptions of different pictures (negative). τ is a temperature parameter controlling the scaling of embedding similarity scores.

D. Product of Experts for Multimodal Concatenation

We adopt the Product of Experts (PoE) approach rather than simple concatenation of modalities, which can rely on spurious features and shortcut solutions [10]. PoE model learns from weaker modalities and balances individual contributions, preventing stronger modalities from dominating the prediction. We begin by processing the features extracted from each unimodality (see Section V-B). First, we concatenate these features from unimodalities and transform them through feed-forward layers (FFN):

$$z_M = \text{FFN}_M([f_S(S_i); f_A(S_i); f_T(T_i); f_I(I_i)]) \quad (2)$$

where $[;]$ denotes concatenations. FFN_M is a feed-forward network to transform multimodal concatenation. z_M denotes the predicted class logits from the multimodal concatenation.

For each unimodality, we also obtain individual predictions of class labels through separate feed-forward networks:

$$z_S = \text{FFN}_S(f_S(S_i)). \quad (3)$$

where FFN is a modality-specific feed-forward network that maps corresponding features to class logits. Similarly, we do this for all other unimodalities.

Next, we aim to combine these predictions efficiently, where the PoE operates in the log space. Rather than performing element-wise multiplication of probabilities, we sum their logarithms values (element-wise addition) because log-space operation is more efficient than probability multiplication.

$$\log z_D = \log z_S + \log z_A + \log z_T + \log z_I. \quad (4)$$

where z_D denotes the final decision logits.

This combination ensures that unimodalities collectively contribute to the final predictions, preventing the reliance on spurious features or a dominated modality. z_D represents a more robust prediction than multimodal concatenation, since it implements dynamic sample weighting: each modality's contribution is adaptively adjusted based on its confidence for each input example. The final prediction logits are used to compute the cross-entropy (CE) loss for binary classification between NC and MCI.

E. Training

The total loss function is a combination of binary cross-entropy loss ($\mathcal{L}_{\text{CE}}$) and the supervised contrastive learning loss ($\mathcal{L}_{\text{sup\_cont}}$), where $\lambda = 1$ balances the two component weights:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{sup\_cont}} \quad (5)$$

VI. Experiments

A. Baseline Details

These baselines focus on each unimodality: z_S represents speech features extracted from the pre-trained speech encoder. z_A represents handcrafted acoustic features from the feature extractor. z_T represents linguistic features from the text encoder. We include $z_M = z_S + z_A + z_T$ to represent speech and text modality concatenation.

B. Implementation Details

To ensure robust evaluation given small and imbalanced data, we trained our framework with stratified k-fold cross-validation (k=10). Hyperparameters include a learning rate of 1e-5, a batch size of 16, L2 regularization penalty= 0.01, and the Adam optimizer. The framework is trained for 10 epochs and implemented in PyTorch on an A100 GPU.

C. Evaluation Metrics

We adhere to evaluation metrics required by the challenge for binary classification: UAR and F1 score: UAR $= \frac{\sigma + \rho}{2}$ and F1 $= \frac{2\pi\rho}{\pi + \rho}$, where $\sigma = \frac{\text{TN}}{\text{TN+FP}}$ represents specificity, $\rho = \frac{\text{TP}}{\text{TP+FN}}$ represents sensitivity, and $\pi = \frac{\text{TP}}{\text{TP+FP}}$ represents precision. TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

VII. Results and Discussion

We focus on the effect of adding supervised contrastive learning (CL), image embeddings (IE), and PoE into baseline models, which are discussed in section VI-A. We evaluate our framework on (1) the entire dataset ("Both") and further break down the dataset into (2) the language subgroup: English ("En") and Chinese ("Zh"), and (3) the gender subgroup: Female ("F") and Male ("M"). It is particularly noted that z_A represents the handcrafted features and is non-learnable with contrastive learning.

A. Baseline Performance

The baseline details are discussed in VI-A. Shown in Table II, the baseline models use unimodal features. On the combined dataset "Both", speech encoder (z_S) achieves UAR of 69.9%. Acoustic features (z_A) achieve UAR of 68.2%, and the text encoder (z_T) achieves UAR of 68.1%. These results indicate that relying on individual modalities is insufficient for achieving highly accurate MCI detection in the multilingual setting. Regarding subgroups of language and gender, Chinese performance is generally higher than English, and male group is better than the female group.

TABLE II
EVALUATION ACROSS EACH UNIMODALITY WITH CL AND IE. RESULTS INCLUDE ENTIRE DATA AND LANGUAGE AND GENDER SUBGROUP PERFORMANCES. GREEN HIGHLIGHTS VALUES ABOVE THE BASELINE IN EACH COLUMN.

Method	UAR					F1				
	Both	En	Zh	M	F	Both	En	Zh	M	F
z_S	69.9	55.9	74.2	80.3	63.3	80.6	79.6	75.7	84.3	77.5
+ CL	68.0	55.0	74.5	75.5	62.9	80.7	79.2	78.6	82.3	78.5
+ CL + IE	69.0	55.0	72.7	78.8	63.0	80.2	79.2	74.3	83.3	77.2
z_A	68.2	55.0	70.3	77.3	61.6	78.9	79.2	71.8	79.2	75.6
+ IE	68.3	49.5	70.4	78.9	61.3	78.7	78.7	71.6	79.7	75.1
z_T	68.1	55.0	74.7	76.1	62.5	80.6	79.2	79.0	82.8	78.1
+ CL	71.8	60.9	74.6	74.2	71.3	81.5	80.4	72.8	81.1	81.3
+ CL + IE	73.1	61.6	74.1	76.5	71.8	82.0	81.1	71.6	82.6	81.0
z_M	72.4	58.0	83.4	83.1	70.0	81.3	80.3	78.7	87.0	78.4
+ CL	72.6	56.1	73.6	78.0	67.9	82.2	81.5	74.3	86.8	79.5
+ CL + IE	73.5	55.1	77.3	82.5	68.3	83.3	81.0	77.4	85.8	79.6

B. After Adding CL and IE to Unimodality

CL and IE are more effective for the text modality than for the speech modality. As shown in Table II, adding CL to z_T increases UAR from 68.1% baseline to 71.8% (+3.7%). The addition of IE further improves it by 5.0% (73.0%). The F1 score improves from 80.6% to 82.0% after adding both CL and IE. In contrast, speech models show no substantial improvements. These indicate that CL can effectively enhance the model's ability to discriminate linguistic features and semantic differences from multi-picture descriptions. The inclusion of IE provides more picture context in addition to text.

Language and gender subgroups with unimodality. In terms of language-specific performance, CL and IE demonstrate more improvements in English compared to Chinese within the text modality: adding CL and IE noticeably increases the UAR in English from 55.0% to 61.6% (+6.6%). In terms of gender subgroup performance, improvements are consistent, where the text modality with both CL and IE improves the UAR of the female by +9.3% (62.5% to 71.8%).

C. After Combining All Modalities

CE and IE integration with PoE improves multimodal baseline performance. Shown in Table III, our multimodal baseline Z_M achieves UAR of 72.4%, outperforming all unimodal approaches. Incorporating PoE further improves UAR to 73.6%, since PoE encourages individual modalities to contribute collectively to make the final decision. This process aims to enhance model generalization. When we combine PoE, CL and IE, we observe additional improvements compared with Z_M: increases in UAR by +2.8% (72.4% to 75.2%) and F1 score by +2.2% (81.3% to 83.5%). These results imply that the combination of CL and IE with PoE leads to an improved framework for MCI detection.

TABLE III
COMPARISON OF MULTIMODAL BASELINE (z_M), z_M WITH PoE, AND THEIR COMBINATIONS WITH CL AND IE. BOLD VALUES ARE WHERE z_M + PoE OUTPERFORMS z_M. GREEN HIGHLIGHTS INDICATE IMPROVEMENTS ABOVE z_M + PoE.

Method	UAR					F1				
	Both	En	Zh	M	F	Both	En	Zh	M	F
z_M	72.4	58.0	83.4	83.1	70.0	81.3	80.3	78.7	87.0	78.4
z_M + PoE	**73.6**	**58.9**	78.7	78.5	**73.2**	**81.7**	**80.6**	**79.5**	80.9	**82.1**
+ CL	73.8	59.0	78.7	78.8	73.2	82.3	81.3	79.8	81.3	82.0
+ CL + IE	75.2	60.1	78.3	78.6	69.3	83.5	80.6	79.7	84.4	80.2

Language and gender subgroup with multimodality. Language-specific results show that PoE improves English performance, with +0.9% increase in UAR and with a further increase of +2.1% (58.0% to 60.1%) after adding CL and IE. While Chinese shows stronger baseline performance, its UAR decreases after adding PoE. This pattern suggests that PoE balances the integration of modalities across languages, reducing inter-group disparities. Similar results emerge across gender subgroups: the lower-performing female baseline shows improvements of +3.2% in UAR (70.0% to 73.2%) and +3.7% in F1 score (78.4% to 82.1%), while male performance drops, narrowing the gender disparity. Thus, we fairly argue that PoE has the capability to balance learning from different modalities regardless of differences between subgroups for a better generalization. The addition of CL and IE further improves the male F1 score by +3.5% (80.9% to 84.4%), though the female performance remains stable.

PoE reduces disparity across subgroups. The inclusion of PoE leads to improvements either in UAR or F1 score compared to the underperforming baseline subgroups. In the
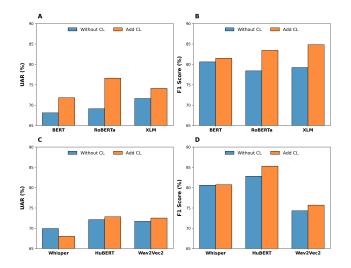

Fig. 2. Performance impact of contrastive learning on different models. **A-B** UAR and F1 scores for linguistic models (BERT, RoBERTa, XLM) with and without CL. **C-D** UAR and F1 scores for speech models (Whisper, HuBERT, Wav2Vec2) with and without CL.

gender subgroup, female has UAR increase of +3.2% (70.0% to 73.2%) and F1 score increase of +3.7% (78.4% to 82.1%). The male performance significantly reduces, thereby reducing the gender disparity. Similarly, since Chinese UAR reduces while English remains stable, narrowing the gap between the previously stronger Chinese baseline. These results exhibit that PoE can effectively elevate worse-performing subgroups while maintaining outcomes from better-performing subgroups.

D. The Impact of Contrastive Learning on Other Models

To further understand the effectiveness of adding CL, we evaluate it across multiple linguistic and speech feature extraction models. Linguistic models include BERT, RoBERTa[5] [23], and XLM[6] [24]. Speech models include Whisper, Hu-BERT[7] [25], and Wav2Vec2[8] [26].

Linguistic models reveal consistent improvements with CL. As shown in Figures 2A and 2B, RoBERTa shows a noticeable improvement, with UAR increasing by +7.5% and F1 score by +5.2%. XLM has consistent improvements (UAR +2.5%, F1 +5.6%). Notably, XLM has a stronger UAR baseline prior to adding CL, likely due to its cross-lingual pre-training. In contrast, the effect of CL on speech features demonstrates limited benefits in Figures 2C and 2D: HuBERT shows a small increase of +0.7% in UAR and +2.5% in F1 score, as well as Wav2Vec2 (UAR +0.8%, F1 score +1.4%). HuBERT has a strong baseline among speech models.

Contrastive learning enhances picture-specific linguistic features more effectively. Our experiments consistently demonstrate that CL better separates linguistic feature embeddings than speech features in high-dimensional spaces. This enhanced separation improves discrimination between

---

[5]https://huggingface.co/FacebookAI/roberta-base
[6]https://huggingface.co/FacebookAI/xlm-roberta-base
[7]https://huggingface.co/facebook/hubert-base-ls960
[8]https://huggingface.co/facebook/wav2vec2-base

descriptions of different pictures in the linguistic domain. However, speech embeddings prove less effective for this separation because they are dominated by speaker-specific characteristics (vocal patterns, speaking style) rather than picture-specific content. Consequently, picture-specific differences remain more distinguishable in linguistic representations than in speech feature space.

VIII. LIMITATION

Our work primarily uses the supervised contrastive learning loss and PoE loss functions. We hypothesize that the results will generalize beyond the loss functions we discussed. Future validation is needed to verify this. Other questions remain open: we explore a limited set of MCI-relevant handcrafted features. Future work will analyze more handcrafted features and how they are correlated with the performance of our framework. We will also compare with other multimodal fusion strategies to compare their performance with ours. In addition, the small and imbalanced data limits the highly performing features and models, which requires future data collection and benchmark efforts.

IX. CONCLUSION

In this paper, we present a framework for MCI detection that addresses the challenges of multilingual and multi-picture analysis. By integrating supervised contrastive learning, image embeddings, and a Product of Experts (PoE) mechanism, our approach enhances MCI-related feature representations while mitigating modality-specific biases. The proposed method achieves notable improvements over the text-only baseline, with a +7.1% gain in UAR (from 68.1% to 75.2%) and a +2.9% gain in F1 score (from 80.6% to 83.5%). Our findings demonstrate that contrastive learning effectively captures picture-specific discriminative features, particularly in the text modality. Furthermore, PoE-based multimodal fusion balances modality contributions and improves generalization by reducing spurious correlations and subgroup disparities.

REFERENCES

[1] G. A. Marshall, M. Dekhtyar, J. M. Bruno, K. Jethwani, R. E. Amariglio, K. A. Johnson, R. A. Sperling, and D. M. Rentz, "The harvard automated phone task: new performance-based activities of daily living tests for early alzheimer's disease," *The journal of prevention of Alzheimer's disease*, vol. 2, no. 4, p. 242, 2015.

[2] Y. Yamada, K. Shinkawa, M. Kobayashi, M. Nishimura, M. Nemoto, E. Tsukada, M. Ota, K. Nemoto, and T. Arai, "Tablet-based automatic assessment for early detection of alzheimer's disease using speech responses to daily life questions," *Frontiers in Digital Health*, vol. 3, p. 653904, 2021.

[3] K. Qi, J. Shi, C. Summerour, J. A. Batsis, and X. Liang, "Exploiting longitudinal speech sessions via voice assistant systems for early detection of cognitive decline," in *HealthCom*. IEEE, 2024, pp. 1–6.

[4] D. Beltrami, G. Gagliardi, R. Rossini Favretti, E. Ghidoni, F. Tamburini, and L. Calzà, "Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?" *Frontiers in aging neuroscience*, vol. 10, p. 369, 2018.

[5] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[6] S. Luz, S. De La Fuente Garcia, F. Haider, D. Fromm, B. MacWhinney, A. Lanzi, Y.-N. Chang, C.-J. Chou, and Y.-C. Liu, "Connected speech-based cognitive assessment in chinese and english," in *Proceedings of Interspeech 2024*, 2024.

[7] J. Chen, J. Ye, F. Tang, and J. Zhou, "Automatic detection of alzheimer's disease using spontaneous speech only," in *Interspeech 2021*, 2021, pp. 2851–2855.

[8] G. Gosztolya and L. Tóth, "Combining acoustic feature sets for detecting mild cognitive impairment in the interspeech'24 taukadial challenge," in *Proc. Interspeech 2024*, 2024, pp. 957–961.

[9] D. Ortiz-Perez, J. Garcia-Rodriguez, and D. Tomás, "Cognitive insights across languages: Enhancing multimodal interview analysis," *arXiv preprint arXiv:2406.07542*, 2024.

[10] R. Cadene, C. Dancette, M. Cord, D. Parikh *et al.*, "Rubi: Reducing unimodal biases for visual question answering," *Advances in neural information processing systems*, vol. 32, 2019.

[11] Y.-L. Liu, R. Feng, J. Yuan, and Z.-H. Ling, "Clever hans effect found in automatic detection of alzheimer's disease through speech," *arXiv preprint arXiv:2406.07410*, 2024.

[12] J. Cheng, M. Elgaar, N. Vakil, and H. Amiri, "Cognivoice: Multimodal and multilingual fusion networks for mild cognitive impairment assessment from spontaneous speech," in *Interspeech 2024*, pp. 4308–4312.

[13] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 18 661–18 673.

[14] M. L. Gauder *et al.*, "Alzheimer disease recognition using speech-based embeddings from pre-trained models," in *Interspeech*, 2021.

[15] A. Balagopalan and J. Novikova, "Comparing acoustic-based approaches for alzheimer's disease detection," in *Interspeech*, pp. 2846–2850.

[16] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.

[17] P. A. Pérez-Toro, T. Arias-Vergara, F. Braun, F. Hönig, C. A. Tobón-Quintero, D. Aguillón, F. Lopera, L. Hincapié-Henao, M. Schuster, K. Riedhammer *et al.*, "Automatic assessment of alzheimer's across three languages using speech and language features," in *Proceedings of Interspeech*, 2023, pp. 1748–1752.

[18] B. Hoang, Y. Pang, H. Dodge, and J. Zhou, "Translingual language markers for cognitive assessment from spontaneous speech," in *Proc. Interspeech 2024*, 2024, pp. 977–981.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[20] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article," *Frontiers in Psychology*, vol. 12, 2021.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[23] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.

[24] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in neural information processing systems*, vol. 32, 2019.

[25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.