MASA: A Novel Multimodal Foundation Model for L2 Speaking Assessment in Picture-description Scenarios

Bi-Cheng Yan, Fu-An Chao, Hong-Yun Lin, Berlin Chen

National Taiwan Normal University, Taipei, Taiwan
{bicheng, fuanchao, berlin}@ntnu.edu.tw

Abstract

Automatic speaking assessment (ASA) manages to quantify the language competence of second language (L2) learners by providing a proficiency score based on their spoken responses. Existing efforts typically employ a neural grader coupled with a set of handcrafted features to gauge the competence of language in L2 learners from multiple facets. Despite their decent efficacy, these methods are limited by a laborious feature engineering process and largely overlook the utilization of scoring rubrics that are presented to human raters in speaking assessment. In light of this, we put forward a novel Multimodal foundation model for ASA, termed MASA, for use in picture-description scenarios. Our approach effectively streamlines the feature engineering process by leveraging the pre-trained encoders of a multimodal foundation model, and emulates the nuanced scoring behaviors of human raters by incorporating scoring rubrics directly into the modeling process. Furthermore, a simple, training-free method is introduced to alleviate the scoring bias in MASA by contrasting the output distributions derived from the multimodal and single-modal inputs. A series of experiments conducted on a picture-description task of the General English Proficiency Test (GEPT) dataset validates the feasibility and superiority of our method in comparison to several cutting-edge baselines.

Keywords: Automatic speaking assessment, multimodal foundation model, debiasing approach

1. Introduction

Spurred by the increasing global demand for foreign language acquisition in both the workforce and academia, there is a pressing need for assessments of language competence (Larry and Norris, 2024). In response, the development of automatic speaking assessment (ASA) systems has garnered increasing attention, figuring prominently in various fields of computer-assisted language learning (Zechner and Evanini, 2019) and large-scale language testing (Singla et al., 2021). ASA systems offer a broad spectrum of applications to mitigate the disparity between the limited number of language instructors and the expanding population of second language (L2) learners. These applications span from low-stakes contexts, such as providing informative feedback for instructors and learners in course placement (Evanini and Wang, 2013), to high-stakes scenarios, such as serving as a reliable reference for professionals in admission testing (Evanini et al., 2017; Chen and Li, 2016).

ASA seeks to quantify language proficiency of L2 learners along multiple dimensions of language competence, including, delivery (e.g., fluency and pronunciation), language use (e.g., vocabulary and grammar), topic development (e.g., content and discourse), and others (Qian et al., 2019; Zechner and Evanini, 2019). One of the de-facto archetypes of ASA is instantiated in picture-description scenarios, where an L2 learner is presented with supplementary visual materials (like a picture or video) along with open-ended questions, and is then instructed to respond based on personal experiences or opinions, as schematically deipcted in Figure 1. A leading strand of research on ASA adopts

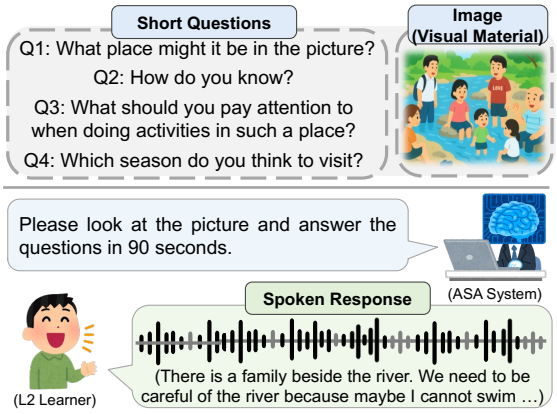

Figure 1: A running example illustrating the speaking assessment in a picture-description scenario, where short prompts paired with an image are presented to L2 learners to elicit authentic spoken response.

presented questions paired with a learner's spoken response to extract a set of proficiency features, which are then fed into a neural grader to predict either a holistic score (i.e., a discrete categorical value for overall speaking proficiency) or analytic scores (i.e., continuous numerical values for specific aspects of language competence). Commonly-used handcrafted features in prior studies for characterizing pronunciation and spoken delivery of L2 learners include confidence scores of recognized linguistic units (e.g., words or phones), time-alignment information (e.g., speaking rate, pause frequency, and filled pauses), and statistical measures of fluency and pitch contours (Zechner et al., 2009; Chen et al., 2010). Beyond spoken delivery, grammatical accuracy and syntactic complexity are derived from the transcriptions of L2 learners'

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5545–5554
11–16 May 2026.

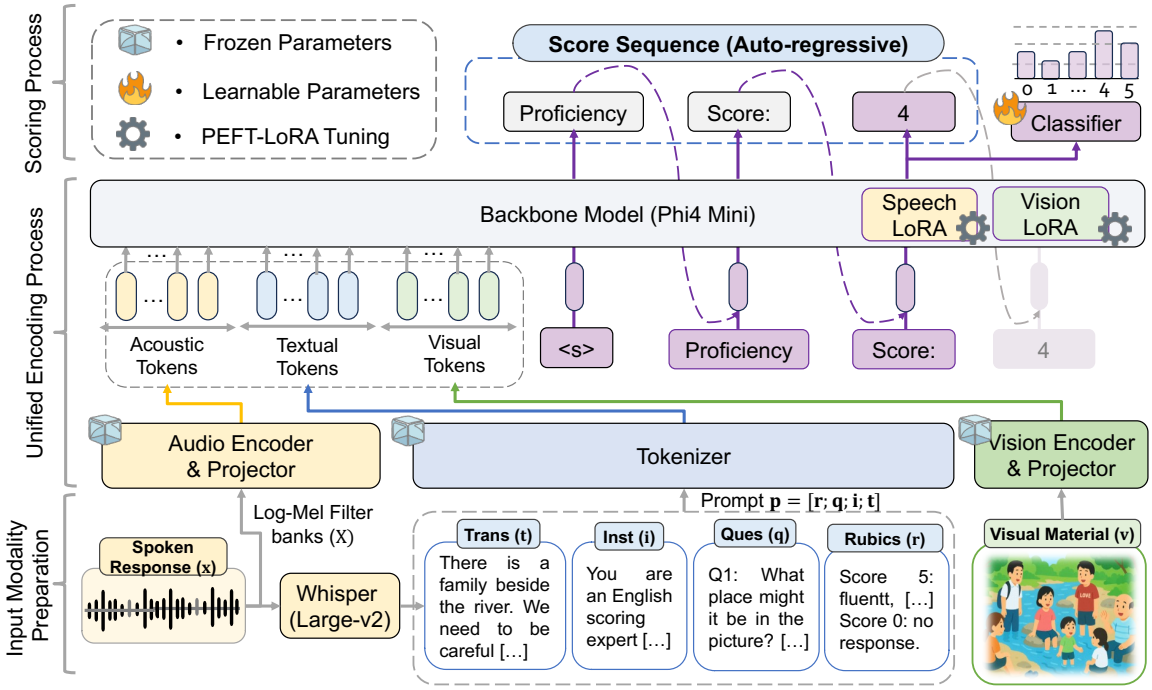

Figure 2: Overview of the proposed MASA architecture, which streamlines the feature engineering process of existing ASA models by leveraging the pre-trained encoders of a multimodal foundation model (e.g., Phi4-Multimodal-Instruction) and incorporates scoring rubrics seamlessly into scoring process. Our MASA first extracts multimodal features from the spoken responses, which are then projected into a unified token space and processed by a language model to generate the proficiency score via both an autoregressive generation process and an auxiliary classifier.

responses with part-of-speech tagging, syntactic dependency parsing, and morphological analysis (Moore et al., 2015; Qian et al., 2018). Lastly, for evaluating content coherence and topical relevance between the learner response and the presented questions, BERT-based semantic features combined with graph-based neural networks are extensively investigated in recent dialogue assessment research (Li et al., 2023; Bannò and Matassoni, 2022; Singla et al., 2022).

Although existing approaches have achieved commendable performance on various speaking assessment tasks by using a set of handcrafted features to capture multiple facets of oral skills, they still suffer from at least two major limitations. First, the lack of a unified cross-modal encoding scheme may make the feature extraction process complicated and error-prone, leading to information loss during the scoring process. Second, prior studies focus exclusively on extracting a set of handcrafted features to gauge language competence; this, however, fails to consider how the extracted features align with the scoring rubrics in an assessment scenario. Building on these observations, we in this paper first propose MASA, a novel Multimodal foundation model for ASA tailored for picture-description scenarios. Our approach streamlines the feature engineering process by leveraging the pre-trained encoders of a multimodal foundation (backbone) model, and mimics the scoring behaviors of human raters by integrating scoring rubrics together with the transcriptions of learners' responses into modeling process. Furthermore, to alleviate the scoring bias and boost the assessment performance, a simple yet effective method is introduced to reduce over-reliance on statistical bias and unimodal prior inherent in the backbone model. Experiments on the picture-description scenario of the General English Proficiency Test (GEPT) dataset demonstrate promising assessment results of MASA, surpassing existing advanced approaches and revealing the unexplored potential of multimodal foundation models for ASA and broader use cases of spoken language understanding.

2. Methodology

2.1 Problem Definition

Given a sequence of questions $\mathbf{q}$ paired with a visual material $\mathbf{v}$, a language learner or test-taker is instructed to produce a spoken response $\mathbf{x}$ that conveys personal experiences and perspectives in relation to $\mathbf{q}$ and $\mathbf{v}$. An ASA model is generally tasked to estimate a categorical proficiency score $y \in \{0, 1, 2, 3, 4, 5\}$ pertaining to $\mathbf{x}$, where each score denotes a distinct level of language competence across the facets of delivery, language use, and topic development, and others. In this paper, the visual material $\mathbf{v}$ and spoken

response $\mathbf{x}$ refer to an image file and a sequence of speech signals, respectively.

2.2 Multimodal Foundation Model for ASA (MASA)

The overall architecture of our MASA architecture is schematically illustracted in Figure 2, where the entire process is executed in three main stages: input modality preparation, unified encoding process, and scoring process. Specifically, after obtaining the multimodal token representations (i.e., acoustic, textual, and visual embeddings), a multimodal backbone language model is employed to fuse the information across all modalities and generate the proficiency score in a natural language expression. In the meantime, an auxiliary classifier utilizes the hidden state to predict the proficiency score, with the aim of enhancing the model's robustness and practical utility. In this paper, our MASA is instantiated through Phi4-Multimodal-Instruct (Abouelenin et al., 2025), which we refer to as Phi4-Multimodal for brevity.

Input Modality Preparation. For each training instance, we first transcribe the spoken response $\mathbf{x}$ into a word sequence $\mathbf{t}$ using a speech recognizer (i.e., Whisper-large-v2). Following this, a prompt template $\mathbf{p}$ is curated to guide the backbone multimodal language model in generating the grading result $\mathbf{y}$, which is a natural language output containing the categorical proficiency score y. The prompt template $\mathbf{p} = [\mathbf{r}; \mathbf{q}; \mathbf{i}; \mathbf{t}]$ comprises four components: the rubric guidelines $\mathbf{r}$, which describe scoring standards used by human experts on a picture-description assessment task; and the instruction $\mathbf{i}$, which specifies the role of the language model and guides it to generate $\mathbf{y}$ for an L2 learner based on the questions $\mathbf{q}$, the transcribed response $\mathbf{t}$, and the information contained in the visual material $\mathbf{v}$.

Unified Encoding Process. Compared to existing ASA models, MASA simplifies the feature extraction pipelines by directly leveraging the pre-trained audio and visual encoders of Phi4-Multimodal. Moreover, by harnessing the strong reasoning capabilities of the backbone multimodal language model, our method can mimic the nuanced scoring behavior of human raters by integrating scoring rubrics in conjunction with the transcription of an L2 learner's response into the scoring process.

A bit of terminology: the prompt $\mathbf{p}$ is first tokenized and mapped into the textual tokens H^t via tokenizer $\mathrm{Enc}_t(\cdot)$ of the backbone language model. Meanwhile, the feature sequences X^a and X^v are extracted from the spoken response $\mathbf{x}$ and the visual material $\mathbf{v}$, respectively, using the pre-trained audio and vision encoders of Phi4-Multimodal, i.e., $\mathrm{Enc}_a(\cdot)$ and $\mathrm{Enc}_v(\cdot)$:

$$\mathrm{X}^a, \mathrm{H}^t, \mathrm{X}^v = \mathrm{Enc}_a(\mathrm{X}), \mathrm{Enc}_t(\mathbf{p}), \mathrm{Enc}_v(\mathbf{v}), \quad (1)$$

where X is a sequence of log-Mel filter-bank features extracted from the speech signal $\mathbf{x}$. The resulting audio and visual features are then projected into a shared token space via modality-specific projectors, which are aligned with the textual modality to enable unified encoding. The unified encoding process then generates a hidden state $\mathbf{e}_n$ at a token position n via the backbone multimodal language model $\mathrm{Phi4}(\cdot)$, formulated as:

$$\mathbf{e}_n = \mathrm{Phi4}(\mathrm{H}^a, \mathrm{H}^t, \mathrm{H}^v, \mathbf{y}_{<n}), \quad (2)$$

$$\mathrm{H}^a, \mathrm{H}^v = \mathrm{Proj}_a(\mathrm{X}^a), \mathrm{Proj}_v(\mathrm{X}^v), \quad (3)$$

where $\mathrm{Proj}_a(\cdot)$ and $\mathrm{Proj}_v(\cdot)$ are audio and vision projectors, and $\mathbf{y}_{<n}$ denotes the sequence of previously generated tokens up to the step $(n-1)$. Through this cross-modal modeling, $\mathbf{e}_n$ is expected to render context-aware interactions, facilitating for understanding and reasoning across multiple input modalities simultaneously.

Scoring Process. To steer the backbone multimodal language model toward proficiency estimation, MASA integrates a lightweight classifier built on top of the backbone language model. This classifier serves as the primary scoring mechanism during inference, offering a direct and efficient way to quantify learner's speaking proficiency. Specifically, in the training phase, the target score y is formatted into a natural language sequence $\mathbf{y}$ with a length of N, which will be utilized in the subsequent scoring process in conjunction with the multimodal language model. For example, as illustrated in Figure 2, the sequence $\mathbf{y}$ is exemplified as:

$$\mathbf{y} = \text{<s> The proficiency score: 4 </s>} \quad (4)$$

The classifier then operates on the hidden state $\mathbf{e}_{N-1}$ to generate the proficiency score via the language model:

$$\hat{y} = \mathcal{P}(\hat{y}|\mathbf{x}, \mathbf{v}, \mathbf{p}) = \mathrm{Softmax}(\mathrm{W}\mathbf{e}_{n-1} + \mathbf{b}), \quad (5)$$

$$\mathbf{e}_{N-1} = \mathrm{Phi4}(\mathrm{H}^a, \mathrm{H}^t, \mathrm{H}^v, \mathbf{y}_{<n-2}). \quad (6)$$

For the training objective of MASA, we first consider a classification loss and an autoregressive generation loss. The overall loss is defined as:

$$\mathcal{L}_{\mathrm{MASA}} = \mathcal{L}_{\mathrm{score-clf}} + \alpha\mathcal{L}_{\mathrm{score-ar}}. \quad (7)$$

Here, $\alpha \in [0,1]$ is a weighting coefficient that balances the two objectives. $\mathcal{L}_{\mathrm{score-clf}}$ is the cross-entropy loss between the predicted distribution and the ground-truth score y while $\mathcal{L}_{\mathrm{score-ar}}$ is the negative log-likelihood of the target sequence $\mathbf{y}$ conditioned on the input modalities:

$$\mathcal{L}_{\mathrm{score-clf}} = -\sum_{c=0}^{5} y_c \log \hat{y}_c, \quad (8)$$

$$\mathcal{L}_{\mathrm{score-ar}} = -\log \mathcal{P}(\mathbf{y}|\mathbf{x}, \mathbf{v}, \mathbf{p}). \quad (9)$$

Furthermore, we keep the entire Phi4-Multimodal frozen during training and only fine-tune the speech and vision LoRA modules. This parameter-efficient tuning strategy allows MASA

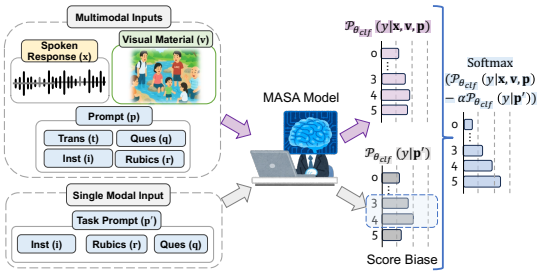

Figure 3: An Illustration of the proposed Single-modal Contrastive Calibration (SCC), where score bias (represented by gray bar charts) induced by single-modal input is mitigated through contrasting with the output distribution derived from the full multimodal input.

to adapt a multimodal foundation model to the speaking assessment task while maintaining its generalization capabilities when only with low-resource L2 ASA datasets.

2.3 Mitigating Scoring Biasing in MASA

Existing literature in natural language processing and computer vision reveals that the hallucination phenomenon stems from insufficient contextual grounding of input modalities (Huang et al., 2025). Multimodal foundation models, in particular, tend to over-prioritize language patterns and rely heavily on partially generated tokens, sidelining the information provided by visual and textual modalities (Agarwal et al., 2020; Wu et al., 2022 ; Leng et al., 2024). Recognizing such a limitation, in this paper, we propose a simple yet effective debiasing mechanism for MASA, dubbed single-modal contrastive calibration (SCC). The proposed SCC aims to rectify the score biases by mitigating two sources of confounding factors, i.e., the over-reliance on inherent language priors from the backbone language model and the statistical biases present in the downstream ASA dataset.

Specifically, given the multimodal input consisting of a spoken response $\mathbf{x}$, a visual material $\mathbf{v}$, and a curated prompt $\mathbf{p}$, the model generates two distinct output distributions. One is conditioned on the full multimodal information $(\mathbf{x}, \mathbf{v}, \mathbf{p})$, while the other is conditioned solely on the single-modal input $\mathbf{p}' = [\mathbf{r}; \mathbf{q}; \mathbf{i}]$, which is a distorted prompt derived by excluding transcribed spoken response from $\mathbf{p}$. Subsequently, SCC computes a calibrated distribution for the proficiency score by contrasting output distributions obtained from the multimodal and single-modal inputs, formulated as follows:

$$\mathcal{P}_{\text{SCC}} = \text{Softmax}\big(\mathcal{P}(\text{y}|\mathbf{x}, \mathbf{v}, \mathbf{p}) - \alpha\mathcal{P}(\text{y}|\mathbf{p}')\big), \quad (10)$$

where a larger α value emphasis on the discrepancy between the multimodal and single-modal output distributions, and $\alpha = 0$ reduces to regular score estimation.

Sets		Ground-truth Score				
		S1	S2	S3	S4	S5
Training		4	61	317	310	27
Development		0	6	39	39	6
Test	Seen-Prompt	0	11	37	39	3
	Unseen-Prompt	0	4	142	137	17

Table 1: Statistics of picture-description section in GEPT Dataset, with each element is the number of spoken responses.

3. Experimental Setups

Picture-description Section in GEPT Dataset. In this paper, a series of experiments were carried out on a dataset curated from General English Proficiency Test (GEPT), a standardized assessment developed by the Language Training and Testing Center (LTTC) in Taiwan[1]. The GEPT targets English learners across all proficiency levels and covers four language skills: listening, reading, writing, and speaking. For our ASA experiments, we focus on the picture-description section of GEPT, in which L2 learners are presented with images and instructed to response the questions. This section comprises 1,199 spoken recordings accompanied by expert-annotated proficiency scores. Each response was independently scored by two qualified experts following the same rubric, and the final categorical proficiency score was obtained by flooring the averaged score. This data was subsequently divided into training, development, and seen-prompt test sets in an 8:1:1 ratio. In addition, an isolated unseen-prompt test set, comprising of 300 recordings, was reserved to evaluate the generalization capability of the developed model. Detailed statistics of the dataset are provided in Table 1.

Implementation Detailed. The proposed MASA method is built upon the Phi4-Multimodal-Instruct[2] model, a lightweight multimodal foundation model that seamlessly integrates speech, vision, and text processing into a unified architecture with up to 5.6B parameters. In the Phi4-Multimodal-Instruct, the audio encoder is composed of 3 convolutional layers with a subsampling rate of 8, followed by 24 Conformer blocks. Each Conformer block is configured with 1024 attention dimensions, 1536 feed-forward dimensions, and 16 attention heads. Furthermore, the vision

[1] The GEPT dataset is subject to restricted access. Please contact the Language Training and Testing Center (LTTC) for licensing and data usage inquiries.

[2] https://huggingface.co/microsoft/Phi-4-multimodal-instruct

Model	Seen-prompt Test			Unseen-prompt Test		
	Accuracy	Weighted F1-Score	Macro F1-Score	Accuracy	Weighted F1-Score	Macro F1-Score
[+]Qian2019	55.56	54.80	-	64.33	62.80	-
[+]SAMAD	65.56	64.80	-	69.67	68.40	-
BERT-ASA	60.44 (±3.20)	59.41 (±3.30)	48.28 (±3.85)	67.00 (±1.39)	65.14 (±0.99)	38.56 (±3.20)
Wav2vec-ASA	52.00 (±2.88)	48.41 (±4.70)	31.72 (±4.85)	52.33 (±2.68)	49.42 (±2.81)	26.89 (±1.35)
Multi-ASA	63.56 (±2.88)	62.85 (±2.92)	44.60 (±5.15)	67.00 (±5.15)	65.37 (±4.65)	41.24 (±4.49)
MASA-AR	66.30 (±3.27)	64.78 (±3.73)	48.38 (±13.19)	69.33 (±2.19)	66.90 (±2.62)	43.26 (±1.98)
MASA	**74.07 (±4.62)**	**72.39 (±6.18)**	**48.66 (±7.95)**	**73.11 (±1.58)**	**71.53 (±2.37)**	**45.20 (±8.91)**
w/o Audio	72.96 (±3.57)	70.83 (±5.24)	47.85 (±13.79)	72.11 (±1.01)	70.25 (±1.29)	43.26 (±5.90)
w/o Trans	71.59 (±3.31)	69.84 (±2.71)	42.05 (±2.58)	71.78 (±2.92)	69.98 (±2.16)	40.46 (±4.34)

Table 2: Performance comparison with several cutting-edge methods on the GEPT dataset in terms of Accuracy (%), Weighted F1-score (%), and Macro F1-score (%). [+]Results for Qian2019 and SAMAD are reported in Peng et al. (2024).

encoder of Phi4-Multimodal-Instruct is based on the SigLIP-400M vision–language model (Zhai et al., 2023), which is further fine-tuned using the LLM2CLIP framework (Huang et al., 2024) on large-scale image-text pairs with an input resolution of 448×448.

Training Configurations. We employed the Adam optimizer with a learning rate of 5e-4 and a batch size of 1. A warm-up strategy was applied, starting at 1/100 of the maximum learning rate and adjusted using a cosine scheduler. To efficiently train the large-scale model architecture, we adopted low-rank adaptation (LoRA) (Hu et al., 2022) and FlashAttention (Dao, 2024), along with low-precision training in bfloat16 to reduce memory footprint and accelerate computation. All experiments were executed on a single NVIDIA RTX 4090 GPU.

Evaluation Metric. 1) Classification Accuracy (%) on proficiency score prediction. 2) Macro F1-score (%) measures the unweighted average of the F1-scores across all proficiency scores, treating each score equally regardless of its frequency. 3) Weighted F1-score (%) computes a class-frequency-weighted average of F1-scores, where each class is weighted by its number of true instances (i.e., class support). Weighted F1-score is important to multi-class classification tasks with datasets having imbalanced or skewed labels, as it provides a balanced evaluation.

To mitigate the effects of randomness, we conducted 3 independent training trials, each initialized with a distinct random seed and trained for 10 epochs. For all evaluation metrics, we reported the mean and standard deviation computed across the best-performing epochs from each trial, where the best epoch was determined by the minimum cross-entropy loss observed on the development set.

Comparative Methods. The proposed MASA method will be benchmarked against top-of-the-line methods from three major categories of ASA techniques. 1) feature-engineering based methods: **Qian2019** is an iconic ASA model that extracts a diverse set of handcrafted features pertaining to fluency, pronunciation, and syntactic complexity based on the transcriptions of learners' spoken responses and the time-alignment information derived from their speech signals (Qian et al., 2019). Subsequently, **SAMAD** advances the neural architecture of Qian2019 by incorporating Transformer blocks and introduces a soft-label optimization to mitigate the label imbalance issues in ASA (Peng et al., 2024). 2) Self-supervised learning-based methods: **Wav2vec-ASA** mitigates the data scarcity problem in L2 English data by resorting to a large-scale pretrained audio encoder (i.e., Wav2vec2.0-

XLSR-300M[3]) to extract acoustic representations from learners' speech, followed by proficiency score prediction via a lightweight linear classifier. In contrast, **BERT-ASA** relies on textual features extracted from a pre-trained text encoder (i.e., BERT-base-uncased[4]) for speaking assessment based on the transcriptions of learners' speech. This approach implicitly captures language proficiency via indicators such as grammatical errors, syntactic structures, and lexical choice. Afterwards, **Multi-ASA** combines Wav2vec and BERT models to extract delivery and content features, while incorporating multiple language-use features for speaking assessment. 3) The variants of MASA: **MASA-AR** adopts Phi4-Multimodal as its backbone model and estimates proficiency features through an autoregressive generation process (Lin et al., 2025). Furthermore, two ablated variants of MASA are considered, i.e., **MASA w/o Audio** (the audio signal $\mathbf{x}$ is omitted), and **MASA w/o Trans** (the transcription $\mathbf{t}$ is excluded).

4. Experimental Results

Main Results. At the outset, we compare our MASA with several cutting-edge ASA methods on the GEPT picture-description dataset. From the results in Table 2, we have the following observations. 1) MASA consistently outperforms all baseline models across seen- and unseen-prompt test sets, achieving superior results on all evaluation metrics. This result highlights the effectiveness of leveraging a multimodal foundation model as a promising avenue for ASA. Furthermore, compared with MASA-AR, our MASA obtains a substantial performance gain by steering the backbone language model toward proficiency estimation via a simple linear layer. To further assess the contribution of each modality within MASA, we compare the two ablated variants (viz. MASA w/o Audio and MASA w/o Trans). The corresponding results manifest that the exclusion of transcription deteriorates performance significantly more than the removal of audio signals, underscoring the critical role of textual information in the scoring process. 2) In comparison of self-supervised learning-based methods, we can observe that BERT-ASA achieves better results than the Wav2vec-ASA. This finding once again underscores that content coherence exerts a greater influence than spoken delivery in speaking assessment, in line with the observations of our ablation studies and prior research (Qian et al., 2019; Bannò and Matassoni 2022). In addition, BERT-ASA outperforms Qian2019 with significant margins across all evaluation metrics, suggesting that the large-scale pretrained text encoder effectively internalizes a broad range of linguistic knowledge, including grammatical correctness, syntactic patterns, and lexical variation. This capability not only brings substantial benefits for proficiency assessment but also strengthens the use of pretrained encoders from a multimodal foundation model in ASA. Subsequently, Multi-ASA integrates self-supervised learning features extracted from pretrained speech and text encoders, along with language-use features, into a unified neural architecture. The corresponding results achieve standout performance compared to each individual module (i.e., Wav2vec-ASA and BERT-ASA), highlighting the benefits of multimodal feature integration in ASA. In contrast to Multi-ASA, which merely combines multimodal features using a simple projection layer, MASA performs a unified encoding process to jointly embeds acoustic and textual information into a shared latent space of the backbone language model. This process enables MASA to generate a more holistic representation of L2 learners and to more effectively capture cross-modal interactions. 3) As a previous state-of-the-art ASA model, SAMAD relies on a set of handcrafted features to portray learners' language proficiency across multiple facets, outperforming several self-supervised learning-based methods. Compared with SAMAD, MASA achieves a substantial performance gain by directly modeling multimodal inputs from L2 learners. This gain stems from its unified multimodal encoding strategy and reduced reliance on manual feature engineering, which together allow the model to capture discriminative features from raw data while minimizing information loss during scoring.

Settings	Seen-prompt Test		Unseen-prompt Test	
	Acc	wF1	Acc	wF1
MASA	70.00	**73.17**	**73.11**	71.53
+SCC	**72.22**	70.82	71.33	**72.17**
MASA-HF	83.33	84.01	72.67	73.62
+SCC	**84.44**	**84.11**	**73.00**	**73.96**

Table 3: Evaluation of MASA extensions that combine the proposed Single-modal Contrastive Calibration (SCC) and integrate handcrafted features into the modeling process (MASA-HF). This table reports classification accuracy (Acc, %) and weighted F1-score (wF1, %) to assess the effectiveness of MASA models.

Extension of MASA Models and the Effectiveness of Score Calibration Mechanism. This set of experiments investigates the impact of integrating handcrafted features into MASA and evaluates the effectiveness of the proposed

[3]https://huggingface.co/facebook/wav2vec2-xls-r-300m

[4]https://huggingface.co/google-bert/bert-base-uncased

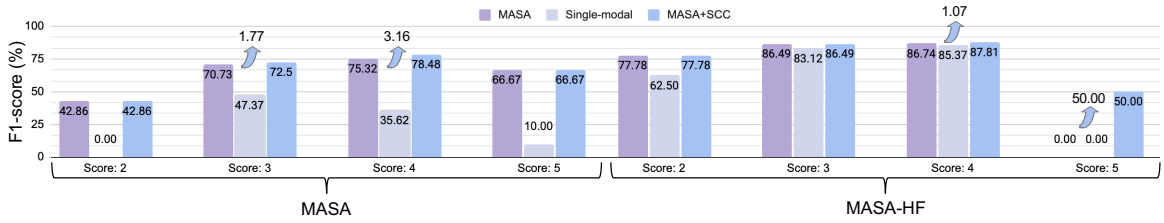

Figure 5: Effectiveness of the proposed Single-modal Contrastive Calibration (SCC), illustrated through F1-score comparisons on the seen-prompt test set for MASA and the MASA model with handcrafted features (MASA-HF) across different proficiency levels.

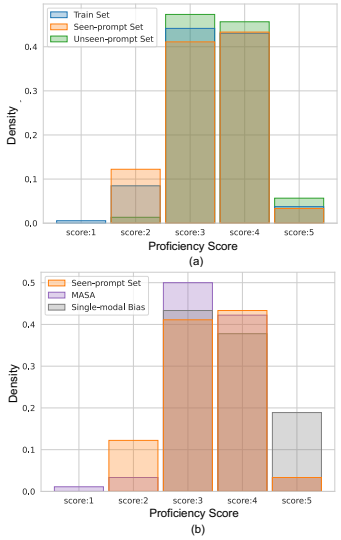

Figure 4: Distribution plots illustrating (a) the data distributions of the training and test sets, and (b) the seen-prompt test set alongside the prediction distributions produced by MASA and its single-modal counterpart.

debiasing approach, single-modal contrastive calibration (SCC). Since SCC is a training-free method, we directly apply it to the MASA model MASA model and its handcrafted feature variant (MASA-HF[5]). Both models are selected from the 3-fold cross-validation experiments and report their SCC-calibrated performance in Table 3. From this table, we observe that MASA-HF significantly outperforms MASA on both the seen- and unseen-prompt test sets, achieving a notable performance margin. This improvement indicates that our proposed model and handcrafted features are highly complementary. Regarding the debiasing mechanism, SCC enhances the performance of MASA models on the seen-prompt test set; however, its generalization to unseen-prompt test set remains limited, yielding only marginal gains or a slight decline in performance.

Exploring the Scoring Bias Inherited in MASA. To better elucidate the varying effects of SCC on the MASA models, we first visualize the score distributions of the training and testing sets in Figure 4(a). We then compare the score distributions of the seen-prompt test set and the model predictions from MASA and its single-modality counterpart in Figure 4(b). A closer inspection of Figure 4(a) shows that the training set (blue) and the unseen-prompt test set (green) exhibit almost identical score distributions, whereas a clear discrepancy appears between the training set (blue) and the seen-prompt test set (orange). Drawing from these distribution characteristics, we postulate that the distributional disparity is crucial for the efficacy of SCC. Specifically, SCC has a limited effect when applied to a distribution nearly identical to the training data, since the prediction bias is minimal. Conversely, the substantial discrepancy in the seen-prompt test set provides greater scope for SCC to adjust and refine the model's predictions.

To further investigate the scoring bias in the backbone language model, Figure 4(b) presents score distributions for the seen-prompt test set (orange) alongside the prediction distributions of MASA (purple) and its single-modality counterpart (gray). The results reveal that the scoring bias of MASA is concentrated around proficiency scores 3 to 5, as its single-modality predictions predominantly fall within this range. This pattern reflects the statistical bias inherent in the GEPT dataset, which is further amplified when MASA estimates proficiency scores solely from the scoring rubrics. Furthermore, relative to the ground-truth labels, MASA tends to over-predict at scores 3 and 4. By simply contrasting the output distributions of MASA and its single-modal counterpart, our SCC strategy effectively mitigate this over-prediction tendency.

Finally, we report the F1-scores for each proficiency score predicted by the MASA models and their single-modality counterparts. As illustrated in Figure 5, the F1-scores of the MASA models (viz. MASA and MASA-HF), their single-modality counterparts, and the SCC-calibrated outputs are depicted in purple, gray, and blue, respectively. Focusing on the MASA model, we observe that the statistical bias is concentrated at

[5] The handcrafted features (i.e., content coherence, pronunciation delivery, and language-use) are extracted following Peng et al. (2024). These features are then concatenated with the last hidden representation of Phi4-multimodal and fed into the classifier through a linear projection layer.

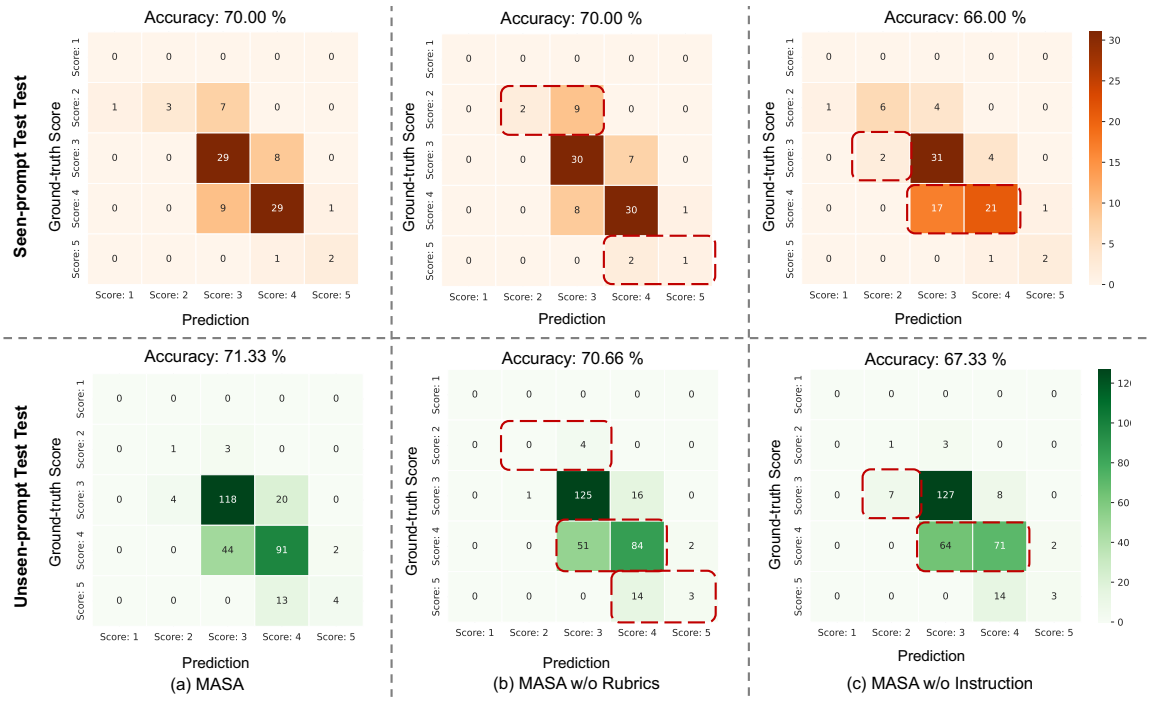

Figure 6: Ablation studies of the MASA models on the test sets, validating the effects of curated prompts that include the scoring rubrics and task instructions. The confusion matrices are shown for (a) the MASA models, (b) the MASA models without scoring rubrics; and (c) the MASA without the task instructions.

proficiency scores 3 and 4, as the predictions derived from single-modality inputs still outperform random guessing. After applying the proposed SCC mechanism, MASA achieve clear performance improvements at these biased scores without sacrificing performance on the remaining ones. This finding demonstrates the effectiveness of SCC in mitigating the systemic bias within MASA. Next, we examine the impact of SCC on MASA-HF. Since the handcrafted features remain uncorrupted in our experiments, MASA-HF with single-modality inputs already exhibits strong performance. Nevertheless, SCC further improves MASA-HF with a moderate gain at score 4 and a notable boost at score 5. The performance gains at score 5 can be attributed to SCC effectively mitigating scoring bias and correcting mispredictions from score 4.

Ablation Studies on the Curated Prompt. In our MASA, a multimodal foundation model serves as the neural grader guided by a curated prompt designed to steer the backbone language model towards speaking assessment. In this set of experiments, we examine the impact of prompt components on MASA. The corresponding results are presented in Figure 6, which visualizes the confusion matrices between the ground-truth scores and predicted proficiency scores for both the seen-prompt and unseen-prompt test sets (the first and second rows, respectively). In Figure 6, we compare the performance of the MASA model with the ablated variants: MASA w/o Rubrics (the prompt excluding scoring rubrics) and MASA w/o Instruction (the prompt excluding task instructions). From these subfigures, we find consistent trends across seen- and unseen-prompt tests, which we summarize as follows. 1) The comparison between Figures 6(a) and 6(b) shows that the inclusion of scoring rubrics in the input prompt allows MASA to better differentiate proficiency within score ranges of [2, 3], and [4, 5], as evidenced by the slight accuracy drop when the rubrics are omitted. 2) By comparing Figures 6(a) and 6(c), we find that the instruction component plays a pivotal role in MASA's prompt design. Its removal results in a pronounced decline in accuracy, most notably at proficiency scores 3 and 4.

5. Conclusion

In this paper, we presented MASA, a novel multimodal foundation model for ASA, designed to quantify learners' language competence in the picture-description scenario. MASA simplifies the feature engineering process by leveraging the pre-trained encoders of Phi4-Multimodal and seamlessly incorporates scoring rubrics into the modeling pipeline. To address the scoring bias inherent in the backbone language model, we further proposed a training-free calibration method which rectifies bias by contrasting the output distributions between the multimodal and the single-modal inputs. Empirical results demonstrate that MASA consistently outperforms several state-of-the-art methods on the GEPT picture-description task, showcasing the promise

of multimodal foundation models for ASA. Moreover, the proposed calibration method effectively alleviates scoring bias in MASA, leading to a fairer and more reliable automated speaking assessment.

6. Acknowledgment

This work was supported by the Language Training and TestingCenter (LTTC), Taiwan. Any findings and implications in the paperdo not necessarily reflect those of the sponsor.

7. Limitations and Future Work

The proposed MASA model novelly adopts a multimodal foundation model for speaking assessment, a step that not only opens new research directions for computer-assisted language learning (CALL) but also raises several important questions for future work. The major limitations are discussed in the following.

Limitation in Foundation Model Selection. As an early attempt at employing a multimodal foundation model for ASA, MASA reports empirical results solely based on Phi4-Multimodal-Instruction. This remains an open question how the performance would be affected by adopting other state-of-the-art multimodal foundation models for ASA, such as, Gamma3 (Gemma et al., 2025), Qwen2.5-Omni (Qwen et al., 2024), and Flamingo3 (Goel et al., 2025).

Limitation in the Single-modal Bias. To mitigate the scoring bias in the backbone language model, we propose the single-modal contrastive calibration (SCC) for MASA, which rectifies the predicted scores by contrasting the output distributions derived from multimodal and single-modal inputs. In our preliminary design of SCC, the single-modal input was defined by excluding the transcription, learner's speech, and visual information from the model inputs. This inevitably raises the question of how the scoring bias would be affected when the inputs involve multiple modalities. In future work, we plan to explore more deliberate modality combinations to better characterize and control the scoring bias in MASA.

Limitation in the Practical Application. Speaking assessment in large-scale English proficiency tests typically comprises a variety of task types, including read-aloud, short-dialogue, and picture-description tasks. Although MASA demonstrates promising performance on the picture-description task, deploying it across different test scenarios presents a notable practical limitation. Specifically, the current MASA architecture requires separate Phi4-based models for each task type, resulting in prohibitive computational costs for real-world deployment. Moving forward, we plan to extend capability of MASA to handle multiple speaking tasks via task-specific instructions and scoring rubrics.

Limitation in Accent Diversity. The used GEPT dataset merely contains Mandarin L2 learners, which hinders the generalizability of the proposed model and could become untenable when assessing L2 learners with diverse accents. As a part of future work, we plan to extend our current work to other speaking assessment datasets, such as ICNALE (Ishikawa, 2021) and S&I challenge (Qian et al., 2024).

8. Bibliographical References

Abdelrahman Abouelenin et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. in arXiv preprint arXiv:2503.01743.

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages. 9690–9698.

Stefano Bannò and Marco Matassoni. 2022. Proficiency assessment of L2 spoken English using wav2vec 2.0. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pages 1088–1095.

Lei Chen, Keelan Evanini, and Xie Sun. 2010. Assessment of non-native speech using vowel space characteristics. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pages 139–144.

Nancy F. Chen and Haizhou Li. 2016. Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In Proceedings of International Conference on Learning Representations (ICLR).

Keelan Evanini and Xinhao Wang. 2013. Automated speech scoring for Nonnative middle school students with multiple task types. In Proceedings of INTERSPEECH (Interspeech), pages 2435–2439, 2013.

Keelan Evanini, Maurice Cogan Hauck, Kenji Hakuta. 2017. Approaches to automated scoring of speaking for K–12 English language proficiency assessments. In ETS Research Report Series, pages 1–11, 2017.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In arXiv preprint arXiv:2507.08128.

Team Gemma et al. 2025. Gemma 3 technical report. In arXiv preprint arXiv:2503.19786.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In Proceedings of International Conference on Learning Representations (ICLR).

Lei Huang et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. In ACM Transactions on Information Systems, vol. 43, pages. 1–55.

Weiquan Huang et al. 2024. Llm2clip: Powerful language model unlocks richer visual representation. In arXiv preprint arXiv:2411.04997.

Shin'ichiro Ishikawa. 2011. A new horizon in learnercorpus studies: The aim of ICNALE project.Corpora and language technologies in teaching,learning and research, G. Weir, S. Ishikawa, and K.Poonpon, Eds., pages. 3–11.

Davis Larry and John M. Norris. 2024. Challenges and Innovations in Speaking Assessment. Taylor & Francis.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13872–13882.

Jiun-Ting Li, Tien-Hong Lo, Bi-Cheng Yan, Yung-Chang Hsu and Berlin Chen. 2023. Graph-enhanced Transformer architecture with novel use of CEFR vocabulary profile and filled pauses in automated speaking assessment. In Proceedings of the workshop on Speech and Language Technology in Education (SLaTE), pages 109–113.

Hong-Yun Lin, Tien Hong Lo, Yu Hsuan Fang, Jhen-Ke Lin, Chung-Chun Wang, Hao-Chien Lu, Berlin Chen. 2025. The NTNU system at the S&I Challenge 2025 SLA Open Track. In the Workshop on Speech and Language Technology in Education (SLaTE).

Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. 2015. Incremental dependency parsing and disfluency detection in spoken learner English. In Proceedings of the International Conference on Text, Speech and Dialogue (TSD), pages 470–479.

Wen-Hsuan Peng, Sally Chen, and Berlin Chen. 2024. Enhancing automatic speech assessment leveraging heterogeneous features and soft labels for ordinal classification. In IEEE Spoken Language Technology Workshop (SLT), pages 945–952.

Team Qwen et al. 2024. Qwen2.5 technical report. in arXiv preprint.

Mengjie Qian, Kate Knill, Stefano Banno, Siyuan Tang, Penny Karanasou, Mark J.F. Gales, and Diane Nicholls. 2024. Speak & improve challenge 2025: Tasks and baseline systems. In arXiv preprint arXiv:2412.11985.

Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, and Xinhao Wang. 2019. Neural approaches to automated speech scoring of monologue and dialogue responses. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8112–8116.

Yao Qian, Rutuja Ubale, Matthew Mulholland, Keelan Evanini, and Xinhao Wang. 2018. A prompt-aware neural network approach to content-based scoring of nonnative spontaneous speech. In Proceedings of IEEE Spoken Language Technology Workshop (SLT), pages 979–986.

Yaman Kumar Singla, Avykat Gupta, Shaurya Bagga, Changyou Chen, Balaji Krishnamurthy, Rajiv Ratn Shah. 2021. Speaker-conditioned hierarchical modelling for automated speech scoring. In Proceedings of the ACM international conference on information & knowledge management (CIKM), pages 1681–1691.

Yaman Kumar Singla, Jui Shah, Changyou Chen, and Rajiv Ratn Shah. 2022. What do audio transformers hear? probing their representations for language delivery structure. In Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW), pages. 910–925.

Yike Wu, Yu Zhao, Shiwan Zhao, Ying Zhang, Xiaojie Yuan, Guoqing Zhao, and Ning Jiang. 2022. Overcoming language priors in visual question answering via distinguishing superficially similar instances. In Proceedings of the International Conference on Computational Linguistics (COLING), pages 5721–5729.

Klaus Zechner and Keelan Evanini. 2019. Automated speaking assessment: Using language technologies to score spontaneous speech. Routledge.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication, volume 51, page 883–895, 2009.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11975–11986.