Synergizing Semantic Anchors and Ordinal Smoothed Cross-Entropy for Speech Fluency Classification

Mulati Kahaer[1,2,3], Sirajahmat Ruzmamat[1,2,3], XuDong Pang[1,2,3], Subinuer Maimaitituerxun[1,2,3], Zaokere Kadeer[1,2,3], Abudurexiti Reheman[3], Wenwen Lu[1,2,3], Panpan Zheng[1,2,3,*], Aishan Wumaier[1,2,3,*]

[1]College of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Xinjiang Multimodal Intelligent Processing and Information Security Engineering Technology Research Center, Urumqi, China
[3]Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China

[*]Corresponding authors: 007652@xju.edu.cn, hasan1479@xju.edu.cn

Abstract

Speech fluency is a core indicator of second language proficiency and a critical component of Computer-Assisted Pronunciation Training (CAPT) systems. Accurate assessment requires models to perceive both macroscopic speech flow trends and microscopic local anomalies. However, existing methods struggle to bridge the semantic gap between static expert priors and dynamic temporal representations, while often overlooking the inherent ordinal nature of fluency scores. To address these challenges, we first construct a set of expert features targeting fluency disruptions and rhythmic regularity to provide explicit linguistic priors. Building on this, we propose the Multimodal Multi-Stream Fusion Classification (MMSFC) network. It employs a Mutual Cross-Attention (MCA) mechanism that leverages these expert features as "semantic anchors" to actively guide Whisper's temporal representations and integrate decoder contexts, achieving deep interaction between global priors and local dynamics. Furthermore, we propose the Ordinal Smoothed Cross-Entropy (OSCE) loss. By constructing distance-aware soft target distributions coupled with confidence-adaptive smoothing and boundary enhancement, OSCE explicitly models ordinal relationships to resolve boundary ambiguity. Experiments on SpeechOcean762 show MMSFC achieves 83.40% accuracy, significantly outperforming strong baselines. Notably, OSCE also demonstrates superior generalization potential in cross-domain CV and NLP tasks. Our code is available at https://github.com/speech26ai/MMSFCCode.

1 Introduction

With the continuous advancement of globalization, cross-cultural communication has become increasingly frequent, and mastering a second language (L2) has emerged as an urgent need and a shared goal for many individuals. Automatic speech fluency assessment is a pivotal component of Computer-Assisted Pronunciation Training (CAPT) and is vital for L2 learning (Tejedor-García et al., 2018; Chen and Li, 2016). Unlike phoneme-level pronunciation quality, fluency is a multi-dimensional suprasegmental attribute involving speech rate, pause distribution, and prosodic rhythm (Kallio et al., 2022). Accurate assessment requires models to perceive both macroscopic speech flow trends and capture microscopic local anomalies (e.g., unnatural pauses).

Existing methods generally fall into two categories: hand-crafted statistical feature methods relying on expert knowledge and feature extraction methods based on pre-trained models, each with its own limitations. The former yields features that are interpretable and reflect global priors, but often lose fine-grained temporal information after being aggregated into static vectors. While pre-trained audio models (e.g., Wav2vec and Whisper) excel at capturing complex temporal dependencies (Baevski et al., 2020; Radford et al., 2022), their representations are generally learned in a task-agnostic fashion, without explicitly encoding fluency-relevant linguistic priors. While simple concatenation of these two types of features is common, it fails to achieve deep complementary interaction.

Furthermore, fluency assessment is inherently an ordinal classification task, yet existing methods mostly treat it as an independent classification problem based on standard Cross-Entropy (CE) (Panda et al., 2023; Preciado-Grijalva and Brena, 2018), ignoring the ordinal distance between grades. Standard CE treats all error predictions equally, while conventional Label Smoothing distributes probability uniformly to non-target classes, failing to reflect the ordinal semantic that predictions should approach the ground truth (Kim et al., 2025; Manuel

Findings of the Association for Computational Linguistics: ACL 2026, pages 31018–31029

Vargas et al., 2022). This neglect of grade progressiveness renders the model unable to distinguish between "adjacent errors" and "distant errors", thereby degrading discriminative performance, especially on samples with ambiguous boundaries.

To address these challenges, we propose the Multimodal Multi-Stream Fusion Classification model (MMSFC). To overcome fusion bottlenecks, we design a Mutual Cross-Attention (MCA) mechanism. It utilizes expert-based statistical features as "semantic anchors" to actively query dynamic acoustic representations, guiding the model toward prosodically relevant temporal segments. Additionally, ASR decoder states are integrated to provide linguistic context without extra text encoders. Furthermore, we propose the Ordinal Smoothed Cross-Entropy (OSCE) loss to explicitly model ordinal relationships. By constructing distance-aware soft targets and incorporating confidence-adaptive smoothing with boundary enhancement, OSCE dynamically adjusts regularization based on prediction uncertainty. This ensures the model captures grade progressiveness while robustly handling samples with ambiguous boundaries.

The main contributions of this paper are summarized as follows:

1. We propose MMSFC, a multimodal fusion framework that seamlessly integrates handcrafted expert statistical features, Whisper-based acoustic representations, and decoder semantics. Experiments on SpeechOcean762 show that our method significantly outperforms strong baselines.

2. We design an MCA mechanism that uses the proposed fluency features as guidance to modulate the attention distribution over pre-trained acoustic representations. This effectively bridges the semantic gap between global statistical features and local temporal dynamics, achieving deep feature-level complementarity.

3. We propose the OSCE loss, which models ordinal relations via distance-aware label distributions, adaptive smoothing, and boundary enhancement. It alleviates the issues where standard losses ignore inter-class distances and apply unreasonable uniform smoothing. We further demonstrate its cross-domain generalization on public CV and NLP datasets.

2 Related Work

Research on automatic speech fluency assessment has evolved from expert-knowledge-based feature engineering to end-to-end deep representation learning, and recently to multi-modal feature fusion. Early approaches primarily relied on acoustic and prosodic features defined by phonetics experts. Systems like SpeechRater (Zechner et al., 2009) and others (Bhat et al., 2010; Evanini and Wang, 2013; Preciado-Grijalva and Brena, 2018) utilized ASR alignments to calculate global statistics (e.g., speech rate, silence duration) for regression or classification. Specific studies have further explored the effectiveness of these features in various contexts, such as using specific prosodic indicators for children's reading assessment (Dimzon and Pascual, 2020) or optimizing speech rate quantification via Bayesian modeling (Yazawa and Konishi, 2025). Notably, Huaijin et al. (2023) found that while deep models excel at general fluency, handcrafted features remain superior in detecting specific disfluencies like filled pauses, highlighting their irreplaceable guiding role in capturing global priors.

With the advent of deep learning, research shifted toward automatic feature extraction to capture complex non-linear relationships. CNNs (Chung et al., 2017) and sequence models like BiLSTMs (Fu et al., 2022) were adopted to model long-term dependencies from raw acoustic features. Recent approaches have explored specialized embeddings, such as attentive X-vectors with ordinal-aware loss functions (Sammit et al., 2022) or contrastive acoustic word embeddings (Wang et al., 2024), to improve scoring precision. Furthermore, Self-Supervised Learning (SSL) models like Wav2vec 2.0 have shown great potential in learning general speech representations without text transcriptions (Liu et al., 2023). However, these pure data-driven methods often learn high-dimensional features that lack explicit linguistic guidance and interpretability, leading to potential overfitting on limited data.

To leverage the strengths of both paradigms, feature fusion has become a mainstream direction. Early attempts combined prosodic and lexical features to quantify fluency (Deshmukh et al., 2009). More recent multi-modal frameworks fuse acoustic embeddings with textual semantics via attention mechanisms (Grover et al., 2020; Liu et al., 2022). For acoustic-level fusion, Li et al. (2024) combined

traditional paralinguistic features with fine-tuned Wav2vec 2.0 representations. Nevertheless, most existing works rely on shallow concatenation strategies. As noted in recent studies, simply stacking explicit global features with implicit temporal representations fails to achieve deep complementary interaction, limiting the model's ability to cover the multi-dimensional attributes of fluency comprehensively.

3 Proposed Method

This section details the proposed end-to-end speech fluency assessment model, as shown in Figure 1. The model consists of three key components: (1) Multi-source feature extraction, which comprehensively utilizes hand-crafted fluency features, pre-trained acoustic representations, and linguistic representations from Automatic Speech Recognition (ASR) decoding; (2) An Encoder-side MCA fusion network, which achieves deep fusion of static statistical features and dynamic temporal representations through bidirectional cross-modal interaction; (3) An OSCE loss function, which designs a distance-aware adaptive label smoothing mechanism for the ordinal classification task.

3.1 Multi-source Feature Extraction

3.1.1 Fluency Features

To incorporate explicit linguistic priors, we construct a 12-dimensional set of expert statistical features $\mathbf{f} \in \mathbb{R}^{12}$, divided into two complementary subsets: 1) Fluency Disruption: comprising pseudo-syllable rate, coefficient of variation of segment duration, speech-to-total duration ratio, silent pause rate, mean silent pause duration, and standard deviation of silent pause duration. These metrics directly quantify discontinuities such as pauses and hesitations; 2) Rhythmic Regularity: including proportion of autocorrelation peaks, mean autocorrelation peak prominence, rhythmic consistency (interval stability), energy proportion in the modulation spectrum, peak frequency within the band (approximating syllable rate), and peak amplitude (beat strength). These features capture prosodic periodicity at the syllable level.

These features are strictly alignment-free: silent pauses are detected via energy thresholds, pseudo-syllable rates are estimated by clustering energy envelope peaks, and rhythmic regularity is captured through autocorrelation and modulation spectrum analysis in the 2–8 Hz syllable band. These static features serve as semantic anchors, effectively complementing dynamic deep acoustic representations.

3.1.2 Pre-trained Acoustic Representations

We employ Whisper-large-v3[1] as the backbone acoustic model. Whisper is pre-trained with weak supervision on 680,000 hours of multilingual and noisy data, demonstrating excellent noise robustness and cross-lingual generalization. Given audio input $\mathbf{x}$, the Whisper encoder produces temporal acoustic features $\mathbf{E} = \mathrm{Encoder}(\mathbf{x}) \in \mathbb{R}^{T \times d_\mathrm{e}}$, where $d_\mathrm{e} = 1280$ and T represents the time series length. These features capture rich acoustic-phonetic information, ranging from phoneme-level details to utterance-level prosodic patterns, which are suitable for fluency assessment.

3.1.3 ASR Decoder Representations

In addition to acoustic features, we extract decoder hidden states from Whisper's autoregressive decoding process. During transcription generation, the decoder produces hidden representations $\mathbf{D} = \mathrm{Decoder}(\mathbf{E}) \in \mathbb{R}^{S \times d_\mathrm{d}}$, where $d_\mathrm{d} = 1280$ and S is the token sequence length. These representations encode linguistic and semantic information aligned with the acoustic content, providing complementary text cues for fluency assessment. Unlike using external language models, this approach ensures perfect temporal alignment between acoustic and linguistic representations while maintaining computational efficiency without requiring extra text encoding.

3.2 Model Architecture

The proposed model architecture aims to deeply combine global statistical information with fine-grained temporal information. This leverages their complementary strengths to overcome the limitations of single feature sources. To achieve this, we design a three-stage cascaded processing flow. First, a projection layer generates latent query vectors. Then, a MCA mechanism on the encoder side enables bidirectional interaction between global priors and local acoustics. Finally, semantic context is injected on the decoder side. This design ensures deep fusion and complementarity of multimodal information within a unified latent space.

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

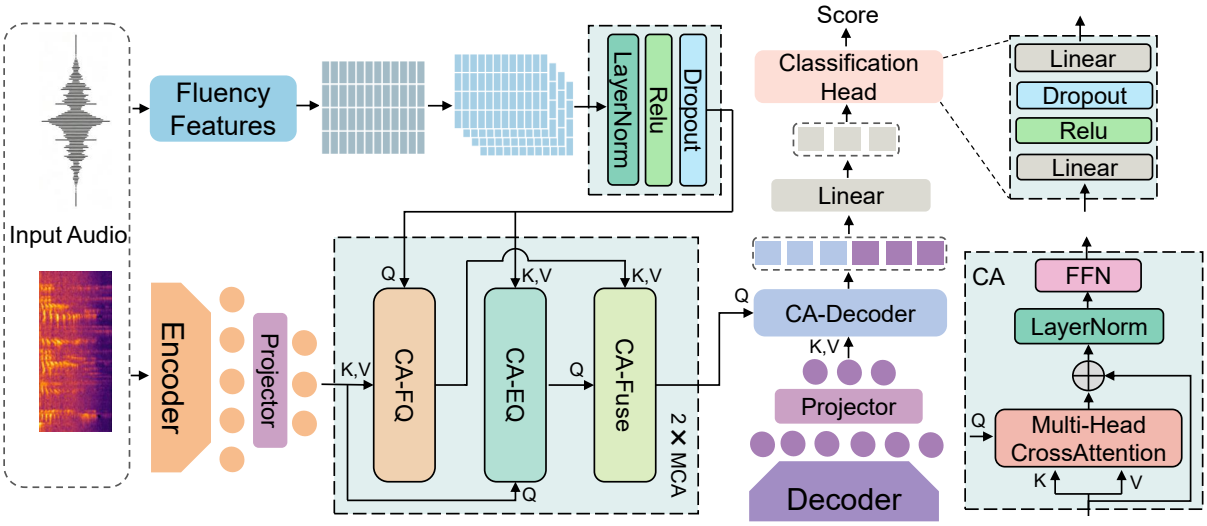

Figure 1: The overall architecture of the MMSFC model.

3.2.1 Projection and Latent Query Generation

Let the temporal acoustic representation from the Whisper encoder be $\mathbf{H}_{\text{enc}} \in \mathbb{R}^{T \times D_{\text{w}}}$, and the handcrafted static fluency feature vector be $\mathbf{f}_{\text{static}} \in \mathbb{R}^{d_{\text{f}}}$. Since they differ significantly in dimension and modality, we first map them to a unified hidden space D_{hidden}.

For temporal features $\mathbf{H}_{\text{enc}}$, we map them to $\tilde{\mathbf{H}}_{\text{enc}} \in \mathbb{R}^{T \times D_{\text{hidden}}}$ using a projection module containing a linear layer, Layer Normalization (LayerNorm), and ReLU activation.

For the static vector $\mathbf{f}_{\text{static}}$, we design an Anchor-based Latent Query Generation mechanism. Specifically, these low-dimensional statistics are mapped to high-dimensional latent vectors serving as semantic anchors, which guide the model towards acoustic segments relevant to specific fluency attributes (e.g., silence patterns, rhythmic strength). Algebraically, we expand $\mathbf{f}_{\text{static}}$ via a learnable projection: $\mathbf{v} = \mathbf{W}_{\text{expand}}\mathbf{f}_{\text{static}} + \mathbf{b}$, where $\mathbf{W}_{\text{expand}} \in \mathbb{R}^{(N_{\text{q}} \cdot D_{\text{hidden}}) \times d_{\text{f}}}$. The vector $\mathbf{v}$ is then reshaped into N_{q} distinct query vectors. To ensure independent semantic representation, we apply Per-Query LayerNorm and an activation function σ: $\mathbf{Q}_{\text{static}} = \sigma(\text{LN}_{\text{per-query}}(\mathbf{v})) \in \mathbb{R}^{N_{\text{q}} \times D_{\text{hidden}}}$. These vectors serve as "semantic anchors" for subsequent interaction stages.

3.2.2 Encoder-Side Interaction: Mutual Cross-Attention

To achieve bidirectional interaction between global priors and local acoustics, the MCA module stacks two interaction layers. We employ the standard Cross-Attention block structure (Vaswani et al., 2017), which comprises Multi-Head Attention and Feed-Forward Networks with residual connections and LayerNorm. For each layer, let $\mathbf{H}_{\text{curr}}$ and $\mathbf{Q}_{\text{curr}}$ denote the input temporal features and semantic anchors.

In the first step, the CA-EQ module performs fluency injection by using $\mathbf{H}_{\text{curr}}$ as Query and $\mathbf{Q}_{\text{curr}}$ as Key/Value, formalized as $\mathbf{H}_{\text{eq}} = \text{Attn}(\mathbf{H}_{\text{curr}}, \mathbf{Q}_{\text{curr}}, \mathbf{Q}_{\text{curr}})$.

Subsequently, the CA-FQ module executes acoustic extraction to allow anchors to actively aggregate relevant segments. Crucially, it queries the original projected acoustics $\tilde{\mathbf{H}}_{\text{enc}}$ to avoid feature drift, calculated as $\mathbf{Q}_{\text{fq}} = \text{Attn}(\mathbf{Q}_{\text{curr}}, \tilde{\mathbf{H}}_{\text{enc}}, \tilde{\mathbf{H}}_{\text{enc}})$.

Finally, the CA-Fuse module integrates the aligned features using $\mathbf{H}_{\text{eq}}$ as Query and $\mathbf{Q}_{\text{fq}}$ as Key/Value to obtain the prosody-aware features: $\mathbf{H}_{\text{fuse}} = \text{Attn}(\mathbf{H}_{\text{eq}}, \mathbf{Q}_{\text{fq}}, \mathbf{Q}_{\text{fq}})$.

3.2.3 Decoder-Side Semantic Integration

To incorporate linguistic context, we fuse the encoder output $\mathbf{H}_{\text{fuse}}$ with Whisper decoder features $\mathbf{H}_{\text{dec}}$. First, we employ a projector (linear transformation) to map the decoder features to the unified hidden space D_{hidden}, yielding $\tilde{\mathbf{H}}_{\text{dec}}$. Then, a cross-attention module uses the prosody-aware features $\mathbf{H}_{\text{fuse}}$ as Query and the projected semantic features $\tilde{\mathbf{H}}_{\text{dec}}$ as Key and Value to retrieve linguistic information, computed as $\mathbf{H}_{\text{cross}} = \text{Attn}(\mathbf{H}_{\text{fuse}}, \tilde{\mathbf{H}}_{\text{dec}}, \tilde{\mathbf{H}}_{\text{dec}})$. Finally, we generate the final fused representation $\mathbf{H}_{\text{final}}$ (i.e., via concatenation with the encoder features and linear projec-

tion), which is fed into the classification head.

The classification head adopts an MLP structure, comprising a dimensionality reduction linear layer ($D_{\text{hidden}} \rightarrow D_{\text{hidden}}/2$), ReLU activation, and Dropout regularization, finally projecting to prediction logits for C fluency levels.

3.3 Ordinal Smoothed Cross-Entropy Loss

Given that the fluency rating task is essentially an ordinal classification problem, standard classification loss functions often ignore the relative distance relationships between grades. Therefore, we propose an OSCE Loss. This loss function explicitly encodes ordinal relationships by constructing a distance-aware soft target distribution. It also introduces a confidence-based adaptive smoothing mechanism for sample-level dynamic regularization. This effectively balances classification accuracy and ordinal consistency during training.

3.3.1 Distance-Aware Label Distribution

To encode the ordinal distance between classes into the target space, we construct a soft target distribution centered on the ground truth label with power-law decay on both sides. For sample i and its ground truth label y_i, we define the unnormalized weight $w_{i,k}$ for class k as the inverse power function of its absolute distance from the ground truth:

$$w_{i,k} = \frac{1}{(|k - y_i| + \delta)^{\lambda}} \tag{1}$$

where λ (default 2.0) controls the decay sharpness, and $\delta > 0$ denotes a small numerical stabilizer (we use $\delta = 10^{-8}$ in all experiments). At this point, the probability mass for non-ground-truth classes ($k \neq y_i$) is no longer uniform but strictly follows a distance-driven power-law distribution.

The final soft target distribution $q_{i,k}$ incorporates the smoothing intensity ϵ_i, defined as:

$$q_{i,k} = \begin{cases} 1 - \epsilon_i, & \text{if } k = y_i \\ \epsilon_i \cdot \frac{w_{i,k}}{\sum_{j \neq y_i} w_{i,j}}, & \text{if } k \neq y_i \end{cases} \tag{2}$$

This distribution ensures that the model optimization focuses not only on classification accuracy but is also constrained by fine-grained ordinal deviations.

3.3.2 Confidence-Adaptive Smoothing

To overcome the limitation of traditional label smoothing which applies a constant penalty to all samples, we propose a confidence-based Dynamic Regularization strategy. This strategy uses the model's posterior probability of the true class $p_{t,i} = p(y_i|\mathbf{x}_i)$ (with gradients truncated) as a proxy for sample learning difficulty. It adaptively adjusts the smoothing intensity ϵ_i for each sample:

$$\epsilon_i = \text{Clamp}\big(\epsilon_{\text{base}} \cdot (1 + \alpha \cdot (1 - p_{t,i})), \epsilon_{\min}, \epsilon_{\max}\big) \tag{3}$$

where ϵ_{base} is the base smoothing rate and α is the adaptive intensity coefficient. The core advantage of this mechanism is its sample adaptability: it automatically increases smoothing on "hard samples" where model prediction is uncertain to suppress overfitting to noise, while reducing smoothing on "easy samples" with high confidence to maintain feature discriminability. This results in a more robust training process.

3.3.3 Boundary Class Enhancement

Considering that the probability distributions of boundary classes (e.g., score 0 and score $C - 1$) in ordinal regression tend to shift towards the center, we introduce a conditional decay factor when constructing $w_{i,k}$. When the ground truth label y_i belongs to a boundary class, we modify the decay factor λ to λ':

$$\lambda' = \begin{cases} \lambda \cdot \beta, & \text{if } y_i \in \{0, C-1\} \\ \lambda, & \text{otherwise} \end{cases} \tag{4}$$

where $\beta > 1$ is an enhancement factor (we set $\beta = 1.3$ in this paper). This operation accelerates the decay rate of the target distribution for boundary samples on non-ground-truth classes. This constructs a more concentrated probability mass, thereby strengthening the model's ability to discriminate extreme scores.

4 Experimental Setup

4.1 Datasets

We evaluate our method on the SpeechOcean762 dataset (Zhang et al., 2021), comprising 5,000 English utterances (approximately 6 hours of audio) from 250 Mandarin-speaking L2 learners. We focus on sentence-level fluency assessment. Following the official protocol, fluency scores are categorized into four ordinal classes based on score ranges: 0–3, 4–5, 6–7, and 8–10. Experiments use the official split with 2,500 utterances for training and 2,500 for testing.

4.2 Implementation Details

All experiments are implemented using the PyTorch framework and trained on a single NVIDIA A40 GPU. We set the batch size to 16 and the maximum number of epochs to 100. Model parameters are updated via the AdamW optimizer, with both the initial learning rate and weight decay coefficient set to 1×10^{-5}. The hidden layer dimension (D_{hidden}) is set to 256. To optimize training convergence, we use the ReduceLROnPlateau scheduler to dynamically adjust the learning rate, combined with early stopping to prevent overfitting. Model performance is evaluated mainly based on Accuracy and weighted F1 score.

4.3 Baselines

To validate the effectiveness of our method, we compare MMSFC with three categories of baselines:

Hand-crafted Feature-based Methods. We evaluate traditional classifiers (SVM, RF, MLP) based on MFCCs, and a recent multimodal fluency assessment model (Liu et al., 2022) (denoted as PSCFluency in this paper) that fuses acoustic, prosodic, and textual features.

Pre-trained Model-based Methods. We employ frozen encoders from Wav2vec 2.0, WavLM (Chen et al., 2022), Whisper, and the audio encoder of CLAP (Wu et al., 2023) (noted for its performance in audio classification) as feature extractors. Additionally, we adapt the SSL+ASR-free model (Liu et al., 2023) by replacing its regression head with a classification head to fit this task.

Advanced Audio Architecture. We introduce MAX-AST (Alex et al., 2024), a state-of-the-art architecture from audio event classification, as the third baseline category. We transfer this generic model to our task as a performance benchmark for speech fluency classification.

5 Results and Analysis

5.1 Main Results

Table 1 presents the comparative results on the SpeechOcean762 test set. Overall, deep learning approaches significantly outperform traditional machine learning methods. Traditional models based on MFCCs stall below 75% accuracy, indicating that relying solely on shallow statistical features is insufficient to capture complex fluency patterns. While PSCFluency improves accuracy to 76.40% via multimodal fusion, it still lags behind pre-trained models with stronger representation capabilities, such as the Whisper Encoder (79.74%).

Model	ACC (%)	F1 (%)
RF	73.24	68.44
MLP	74.84	72.83
SVM	75.20	73.66
PSCFluency	76.40	74.47
CLAP	76.36	75.20
Wav2vec2.0	77.12	75.91
WavLM	79.52	77.14
Whisper	79.74	80.96
SSL+ASR-free	80.72	80.11
MAX-AST	78.08	77.60
MMSFC (Ours)	**82.68**	**82.35**

Table 1: Performance comparison with baseline models on SpeechOcean762.

In the comparison of advanced architectures, MAX-AST underperforms the task-specific SSL+ASR-free method (80.72%). Most notably, our MMSFC model achieves 82.68% accuracy and 82.35% F1 score using only standard Cross-Entropy loss, surpassing SSL+ASR-free by 1.96% in accuracy and 2.24% in F1. This result demonstrates that deep fusion of global static features and local dynamic representations via the MCA mechanism effectively breaks through the performance bottleneck of single-source features.

5.2 Ablation Study: Architecture Effectiveness

To verify the contribution of each component, we conducted a detailed ablation study (see Table 2).

First, the necessity of feature sources is confirmed. Removing fluency features (No-Fluency) drops accuracy to 81.04%, indicating the irreplaceable complementarity between expert priors and self-supervised representations.

Second, the analysis of fusion strategies reveals the hierarchy of deep interaction. Simple feature concatenation (Naive) yields an accuracy of only 81.52%. We find that introducing attention solely at the decoder side while using concatenation at the encoder side (w/o MCA) fails to improve performance (81.52%), suggesting that without feature alignment at the encoder side, subsequent semantic integration is ineffective. Conversely, enabling MCA solely at the encoder side (w/o Dec-Fusion) significantly boosts accuracy to 82.12%. Notably,

Model Variant	Encoder Fusion	Decoder Fusion	ACC (%)	F1 (%)
No-Fluency	—	—	81.04	81.05
No-Decoder	—	—	82.28	81.82
Naive	Concat	Concat	81.52	80.82
w/o MCA	Concat	CA-Decoder	81.52	81.13
w/o Dec-Fusion	MCA	Concat	82.12	81.88
w/o CA-EQ	MCA (no CA-EQ)	CA-Decoder	81.52	80.44
w/o CA-FQ	MCA (no CA-FQ)	CA-Decoder	82.64	81.89
w/o CA-Fuse	MCA (no CA-Fuse)	CA-Decoder	82.00	81.37
MMSFC (Ours)	**MCA**	**CA-Decoder**	**82.68**	**82.35**

Table 2: Ablation study of architecture variants and MCA components.

this variant slightly underperforms the configuration with the decoder completely removed (No-Decoder, 82.28%). This suggests that naive concatenation introduces noise; only through the full attention mechanism (MMSFC) can linguistic semantics be effectively leveraged to localize fluency boundaries, achieving peak performance (82.68%).

Finally, the internal components of MCA are indispensable. Removing CA-EQ, CA-FQ, or CA-Fuse leads to a significant decline in F1 scores. This validates the necessity of the bidirectional interaction mechanism: it is crucial both to inject global priors into local temporal representations to provide macroscopic context and to aggregate specific acoustic details to calibrate statistical features.

5.3 Ablation Study: Loss Function Effectiveness

We compare OSCE with various mainstream loss functions in Table 3.

First, the limitations of baseline methods are evident. Weighted CE and Focal Loss (Lin et al., 2017), designed for imbalance or hard samples, perform worse than the Standard CE. This suggests that in fluency tasks characterized by subjective ambiguity, aggressive re-weighting strategies tend to introduce training noise or overfit edge samples. Similarly, Quadratic Weighted Kappa (QWK) loss (de la Torre et al., 2018) suffers from suboptimal convergence due to gradient instability.

In contrast, OSCE achieves the best performance (83.40% Accuracy). Compared to Label Smoothing (Müller et al., 2019), which assigns uniform probabilities, the core advantage of OSCE lies in its distance-aware mechanism. By constructing a power-law decay distribution based on ground-truth distance, OSCE explicitly encodes the ordinal semantic that "adjacent predictions are preferable to distant ones." Furthermore, the incorporated sample-adaptive smoothing and boundary boost mechanisms dynamically adjust regularization based on uncertainty and reinforce supervision on extreme scores. This effectively mitigates the issue of conservative boundary predictions common in traditional losses, achieving dual improvements in accuracy and ordinal consistency. Quantitatively, this is validated by the Mean Absolute Error (MAE), where OSCE achieves the lowest error (0.1756) compared to Standard CE (0.1824) and QWK (0.1956), confirming its superior ability to penalize distant errors.

Loss Function	ACC (%)	F1 (%)	MAE ↓
Standard CE	82.68	82.35	0.1824
Weighted CE	81.84	81.65	0.1940
Focal	82.36	81.20	0.1916
Label Smoothing	82.52	82.27	0.1844
QWK	81.72	81.16	0.1956
OSCE (Ours)	**83.40**	**83.00**	**0.1756**

Table 3: Performance comparison of different loss functions.

5.4 Hyperparameter Sensitivity and Component Analysis

To determine optimal configurations and dissect internal mechanisms, we first conducted a sensitivity analysis, followed by an ablation study based on the optimal settings.

First, hyperparameter analysis identifies the optimal equilibrium (refer to Figure 2 in Appendix A.1). Performance peaks as the decay factor λ increases to 2.0. Analysis suggests that a smaller λ

leads to an overly flat distribution introducing non-adjacent noise, while an overly large λ degrades the distribution to approximate one-hot vectors, losing ordinal semantics. Furthermore, the optimal base smoothing rate $\epsilon_{\text{base}} = 0.2$ properly allocates probability mass to adjacent classes while maintaining target confidence, effectively achieving a balance between "hard decisions" and "soft supervision."

With the base hyperparameters fixed, we further designed a controlled experiment to dissect the specific contribution of each mechanism (detailed results are shown in Table 4, Appendix A.1). Compared to the baseline (A1), the fixed OSCE (A2) yields improvements in both metrics, proving that ordinal target distributions offer richer supervision signals than One-hot labels. Further decomposition reveals distinct roles: *Adaptive Smoothing* (A3) prioritizes accuracy optimization (83.08%) by dynamically adjusting smoothing intensity, successfully balancing "noise resistance for ambiguous samples" and "confidence for easy samples"; *Boundary Boost* (A4) demonstrates a clear advantage in F1 scores (82.80%) by reinforcing supervision on extreme scores, effectively mitigating the "central tendency bias" (i.e., the tendency to conservatively predict intermediate scores). Ultimately, the full model (A5) combines both to achieve peak performance (Acc 83.40%), revealing a significant synergistic effect between the two mechanisms.

5.5 Cross-Domain Generalization

To verify the generalization potential of OSCE, we extended our evaluation to CV and NLP domains (see Appendix A.2 for dataset details and Table 5 for results).

In the CV task using the Banana Ripeness dataset (Chuquimarca et al., 2023), we employed a pre-trained ResNet101 (He et al., 2016) with a frozen backbone, training only the classification head. This setup tests the optimization capability under constrained conditions with fixed feature extractors. Results show that OSCE outperforms Standard CE. Notably, Label Smoothing degrades accuracy to 85.69%, falling behind the Standard CE baseline (87.27%). This is attributable to the fact that, within a fixed feature space, indiscriminate uniform smoothing further blurs the visual boundaries between ripeness stages, whereas the distance-aware mechanism of OSCE preserves discriminative power.

In the NLP task using the EngSAF dataset (Aggarwal et al., 2025), we utilized the RoBERTa (Liu et al., 2019) model with full fine-tuning. OSCE demonstrates robustness under this setting. Particularly in the challenging zero-shot setting (*Unseen Questions*), OSCE achieves the highest F1 score (51.73%). In contrast, while Focal Loss yields slightly higher accuracy, its lower F1 score suggests a bias towards predicting majority classes; conversely, OSCE proves effective in balancing class predictions.

In summary, whether under feature-frozen or full fine-tuning scenarios, OSCE provides supervision signals with greater semantic value than aggressive re-weighting or uniform smoothing.

6 Conclusion

This paper proposes the MMSFC model, a multimodal fusion framework that takes linguistic and prosodic expert statistical features as semantic anchors. By injecting these global priors into pre-trained acoustic and decoder semantic representations via MCA, the model achieves deep complementarity between static priors and dynamic temporal representations. To address the ordinal classification nature of fluency assessment, we propose the OSCE loss, which explicitly encodes inter-level ordinal relationships through distance-aware soft targets with power-law decay, combined with confidence-adaptive smoothing and boundary class enhancement. This design effectively mitigates boundary ambiguity and central tendency bias.

Experimental results demonstrate that MMSFC significantly outperforms strong baselines on the SpeechOcean762 dataset, while the OSCE loss exhibits excellent cross-domain generalization in both CV and NLP tasks. This study establishes a new paradigm for ordinal-aware speech assessment, with potential extensions to scenarios such as stuttering detection and pathological speech diagnosis. Meanwhile, the proposed OSCE loss provides a general-purpose ordinal learning framework, transferable across CV and NLP domains—applicable to CV ordinal prediction tasks (e.g., medical image grading, image quality assessment) and NLP tasks (e.g., essay scoring, emotion intensity recognition).

Limitations

First, our expert features primarily capture rhythmic disruptions, lacking coverage of complex prosodic nuances like intonation and lexical stress. Future work could introduce finer-grained and more diverse acoustic and prosodic feature sets to

further enhance the model's discriminative capacity for complex fluency patterns. Second, we rely on frozen Whisper representations without task-specific fine-tuning, which prevents the acoustic features from adapting to the target domain and may constrain the performance ceiling. Third, although our signal-based anchors are language-agnostic, current validation is restricted to specific L2 learner populations; further testing on diverse native-language (L1) backgrounds is essential for universal generalization. Finally, while MMSFC is a lightweight and efficient solution, this study lacks a direct comparison with multi-billion-parameter SpeechLLMs, which represents an important future direction for large-scale fluency modeling.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62466058 for the project "Research on Automatic Evaluation Technology of Topic Talk in PSC Test". The authors would also like to thank the reviewers for their valuable comments and suggestions.

Ethics Statement

We confirm that all authors of this study have adhered to the ACL Code of Ethics and the recommended code of conduct. All datasets used in this work are publicly available, and we have cited the sources of all datasets. We think there are no potential risks for this work.

References

Dishank Aggarwal, Pritam Sil, Bhaskaran Raman, and Pushpak Bhattacharyya. 2025. "i understand why I got this grade": Automatic short answer grading (ASAG) with feedback. In *Artificial Intelligence in Education - 26th International Conference, AIED 2025, Palermo, Italy, July 22-26, 2025, Proceedings, Part III*, volume 15879 of *Lecture Notes in Computer Science*, pages 304–318. Springer.

Tony Alex, Sara Ahmed, Armin Mustafa, Muhammad Awais, and Philip JB Jackson. 2024. Max-ast: Combining convolution, local and global self-attentions for audio event classification. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1061–1065.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Suma Bhat, Mark Hasegawa-Johnson, and Richard Sproat. 2010. Automatic fluency assessment by signal-level measurement of spontaneous speech.

Nancy F. Chen and Haizhou Li. 2016. Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–7.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Hoon Chung, Yun Kyung Lee, Sung Joo Lee, and Jeon Gue Park. 2017. Spoken english fluency scoring using convolutional neural networks. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–6.

Luis Chuquimarca, Boris Vintimilla, and Sergio Velastin. 2023. Banana ripeness level classification using a simple cnn model trained with real and synthetic datasets. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 536–543. INSTICC, SciTePress.

Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154. Machine Learning and Applications in Artificial Intelligence.

Om D. Deshmukh, Kundan Kandhway, Ashish Verma, and Kartik Audhkhasi. 2009. Automatic evaluation of spoken english fluency. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4829–4832.

Francis Dimzon and Ronald Pascual. 2020. Computational prosodic features analysis of children's filipino speech for automated oral reading fluency assessment.

Keelan Evanini and Xinhao Wang. 2013. Automated speech scoring for non-native middle school students with multiple task types. In *Interspeech 2013*, pages 2435–2439.

Kaiqi Fu, Shaojun Gao, Xiaohai Tian, Wei Li, and Zejun Ma. 2022. Using fluency representation learned from sequential raw features for improving non-native fluency scoring. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4337–4341. ISCA.

Manraj Singh Grover, Yaman Kumar Singla, Sumit Sarin, Payman Vafaee, Mika Hama, and Rajiv Ratn Shah. 2020. Multi-modal automated speech scoring using attention fusion. *ArXiv*, abs/2005.08182.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Deng Huaijin, Takehito Utsuro, Akio KOBAYASHI, and Hiromitsu Nishizaki. 2023. Comparative evaluation of diverse features in fluency evaluation of spontaneous speech. *IEICE Transactions on Information and Systems*, E106.D:36–45.

Heini Kallio, Antti Suni, and Juraj Šimko. 2022. Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of english with different language backgrounds. *Language and Speech*, 65(3):571–597. PMID: 34479458.

Daehwan Kim, Haejun Chung, and Ikbeom Jang. 2025. Calibration of ordinal regression networks. *Preprint*, arXiv:2410.15658.

Su-Mei Li, Zi-Xi Zhu, Xiao-Ning Li, and Xin-Guang Li. 2024. A method for automatic english oral fluency scoring. In *2024 17th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Jiajun Liu, Huazhen Meng, Yunfei Shen, Linna Zheng, and Aishan Wumaier. 2022. Multimodal automatic speech fluency evaluation method for putonghua proficiency test propositional speaking section. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 260–264.

Wei Liu, Kaiqi Fu, Xiaohai Tian, Shuju Shi, Wei Li, Zejun Ma, and Tan Lee. 2023. An asr-free fluency scoring approach with self-supervised learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Víctor Manuel Vargas, Pedro Antonio Gutiérrez, and César Hervás-Martínez. 2022. Unimodal regularisation based on beta distribution for deep ordinal regression. *Pattern Recognition*, 122:108310.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? *CoRR*, abs/1906.02629.

Ashish Panda, Rajul Acharya, and Sunil Kumar Kopparapu. 2023. Oral fluency classification for speech assessment. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 231–235.

Alan Preciado-Grijalva and Ramon F. Brena. 2018. Speaker fluency level classification using machine learning techniques. *Preprint*, arXiv:1808.10556.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

George Sammit, Zhongjie Wu, Yihao Wang, Zhongdi Wu, Akihito Kamata, Joseph Nese, and Eric C. Larson. 2022. Automated prosody classification for oral reading fluency with quadratic kappa loss and attentive x-vectors. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3613–3617.

Cristian Tejedor-García, Valentín Cardeñoso-Payo, María J. Machuca, David Escudero-Mancebo, Antonio Ríos, and Takuya Kimura. 2018. Improving pronunciation of spanish as a foreign language for l1 japanese speakers with japañol capt tool. In *IberSPEECH 2018*, pages 97–101.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yihao Wang, Zhongdi Wu, Joseph Nese, Akihito Kamata, Vedant Nilabh, and Eric C. Larson. 2024. Improving oral reading fluency assessment through subsequence matching of acoustic word embeddings. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10766–10770.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Kakeru Yazawa and Takayuki Konishi. 2025. A Bayesian Approach to L2 Fluency Ratings by Native and Nonnative Listeners. In *Interspeech 2025*, pages 106–110.

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*, 51(10):883–895. Spoken Language Technology for Education.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An

open-source non-native english speech corpus for pronunciation assessment. In *Interspeech 2021*, pages 3710–3714.

A Appendix

A.1 Hyperparameter Sensitivity and Component Analysis

We present the detailed sensitivity analysis of the decay factor λ and the base smoothing rate ϵ_{base} in Figure 2. Figure 2(a) illustrates the performance trend as λ varies, peaking at $\lambda = 2.0$. Figure 2(b) shows that the optimal base smoothing rate ϵ_{base} is approximately 0.2.

Additionally, Table 4 provides the exact numerical results for the component ablation study discussed in Section 5.4.

A.2 Details of Cross-Domain Generalization

A.2.1 Datasets and Experimental Settings

Computer Vision (CV) Task. We adopt the **Banana Ripeness Images Dataset** for the ordinal classification of banana ripeness. The original dataset contains both real and synthetic images. To ensure evaluation authenticity and difficulty, we only use the *real-image subset*, excluding all synthetic samples. The real subset consists of 3,495 images of Cavendish bananas collected over a complete 28-day ripening cycle, with approximately 150 images captured per day in a laboratory environment under controlled temperature conditions of 15°C–18°C. After quality screening to remove samples with noise, low illumination, occlusion, or abnormal banana positioning, the final dataset is obtained.

Banana ripeness is divided into four ordinal levels: Level A (days 1–6, unripe), Level B (days 7–14, early-ripe), Level C (days 15–22, mid-ripe), and Level D (days 23–28, overripe). The numbers of samples for the four levels are 1,429, 815, 559, and 692, respectively, exhibiting clear class imbalance while maintaining a strict ordinal relationship among categories.

Following the protocol of the original study, the dataset is split into training, validation, and test sets with a ratio of 60%, 20%, and 20%, respectively, while preserving the class distribution across subsets. This task requires the model to predict the ripeness level of a given banana image, which constitutes a four-class ordinal classification problem. We employ a pretrained ResNet101 as the backbone network and freeze its parameters during training, optimizing only the classification head to evaluate the effectiveness of different loss functions under fixed feature representations.

Natural Language Processing (NLP) Task. For the NLP domain, we use the **Engineering Short Answer Feedback (EngSAF) Dataset** for automatic short-answer grading. This dataset contains 119 questions and approximately 5,800 student responses collected from 25 undergraduate and graduate engineering courses, covering diverse subjects such as image processing, water quality management, and operating systems. Each instance consists of a question, a reference answer, a student answer, and the corresponding grading label.

The original score x is mapped to three ordinal grades according to the full score k: responses are labeled as *Correct* when $x = k$, *Partially Correct* when $0 < x < k$, and *Incorrect* when $x = 0$, forming a strict ordinal hierarchy of *Incorrect* $\rightarrow$ *Partially Correct* $\rightarrow$ *Correct*.

The data splits strictly follow the original protocol, and two evaluation scenarios are designed to comprehensively assess generalization performance:

1. **Unseen Answers:** The test set contains new student answers to questions observed during training, evaluating the model's consistency in scoring semantically different responses to the same questions.

2. **Unseen Questions:** All test questions are unseen during training, requiring the model to perform zero-shot grading and directly assessing its cross-question generalization ability.

This task is essentially a three-class ordinal classification problem. Despite substantial differences between the CV and NLP tasks in data modality, sample characteristics, and application scenarios, both exhibit explicit ordinal structures, providing a representative benchmark for systematically evaluating the cross-domain generalization capability of the proposed OSCE loss.

A.2.2 Full Experimental Results

The complete performance comparison of different loss functions on CV and NLP tasks is presented in Table 5. OSCE demonstrates superior generalization capabilities across all settings.

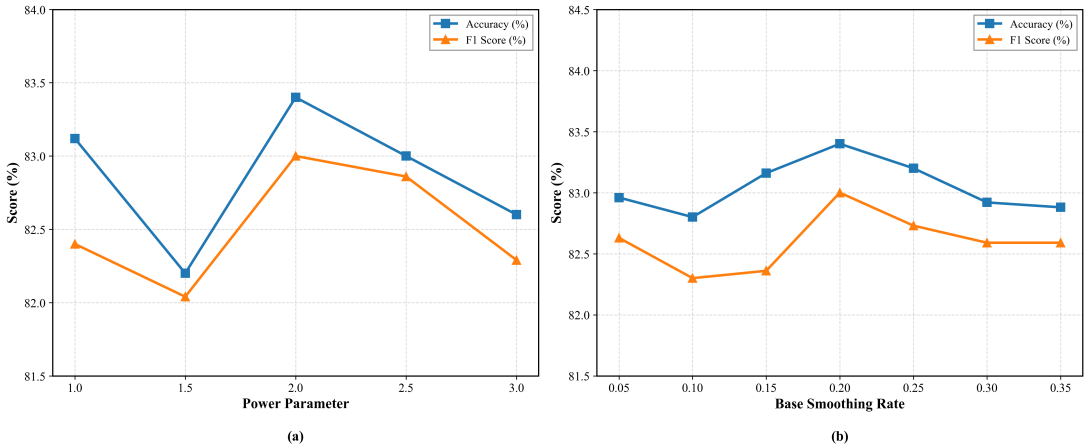

Figure 2: Hyperparameter sensitivity analysis of the OSCE loss function. (a) Impact of Power Parameter λ. (b) Impact of Base Smoothing Rate ϵ_{base}.

ID	Model Configuration	ϵ_{base}	λ	Adaptive	Boundary Boost	ACC (%)	F1 (%)
A1	Standard CE (Baseline)	0	—	×	×	82.68	82.35
A2	OSCE (Fixed)	0.2	2.0	×	×	83.04	82.66
A3	w/ Adaptive Only	0.2	2.0	✓	×	83.08	82.64
A4	w/ Boundary Boost Only	0.2	2.0	×	✓	83.04	82.80
A5	**OSCE (Full)**	**0.2**	**2.0**	✓	✓	**83.40**	**83.00**

Table 4: Component analysis of OSCE with different configurations. "Adaptive" denotes the sample-adaptive smoothing mechanism, and "Boundary Boost" refers to the boundary class enhancement strategy.

Loss Function	Banana (CV)	EngSAF (NLP)	
	Acc / F1 (%)	Ans (Acc / F1, %)	Que (Acc / F1, %)
Standard CE	87.27 / 86.91	76.33 / 76.37	51.63 / 49.30
Weighted CE	88.13 / 87.83	74.59 / 74.57	53.73 / 51.54
Focal	87.84 / 87.91	74.90 / 74.48	54.38 / 51.26
Label Smoothing	85.69 / 85.17	76.12 / 76.13	52.55 / 50.33
QWK	84.84 / 84.16	73.67 / 72.75	54.38 / 51.15
OSCE (Ours)	**88.41 / 88.39**	**77.55 / 77.59**	53.99 / **51.73**

Table 5: Generalization performance on CV and NLP tasks. "Ans" and "Que" denote Unseen Answers and Unseen Questions settings, respectively.