

DPRES: Dual-Perspective Decoupled Representation Learning Framework for Multi-Trait Essay Scoring

Jiong Wang¹, Hao Wu¹, Shengquan Yu^{2,3*}

¹ School of Artificial Intelligence, Beijing Normal University,

² Advanced Innovation Center for Future Education, Beijing Normal University,

³ School of Educational Technology, Faculty of Education, Beijing Normal University
jjoinw@mail.bnu.edu.cn, wuhao@mail.bnu.edu.cn, yusq@bnu.edu.cn

Abstract

Multi-trait automated essay scoring extends holistic scoring by evaluating essays across multiple writing dimensions. In this framework, inter-trait consistency signals provide an effective way to improve prediction across traits. However, learning shared scoring signals may obscure trait-specific features, while enforcing consistency across traits can introduce negative transfer, hindering the model’s ability to distinguish individual traits. To solve these issues, we propose DPRES, a dual-perspective decoupled representation learning framework that disentangles consistency and specificity at both the feature and optimization levels. Specifically, we first propose a residual subspace decomposition strategy that enforces orthogonality between shared scoring features and trait-specific features, thereby disentangling shared scoring signals from trait-specific scoring signals in the latent representation space and enhancing the model’s ability to capture trait-specific scoring features. Based on the subspace decomposition strategy, we further design a consistency–specificity decoupled learning strategy to simultaneously preserve inter-trait correlations and enhance trait-level discrimination. Experiments on public datasets demonstrate that the proposed method consistently outperforms strong baselines.

1 Introduction

By providing feedback across multiple dimensions such as Language, Content, and Organization, multi-trait essay scoring supports a more comprehensive assessment of writing and facilitates learners’ overall writing development. In contrast to holistic scoring approaches (Taghipour and Ng, 2016; Dong et al., 2017; Yang et al., 2020; Elmassry et al., 2025), which only evaluate the overall quality of an essay, multi-trait essay scoring (MES)

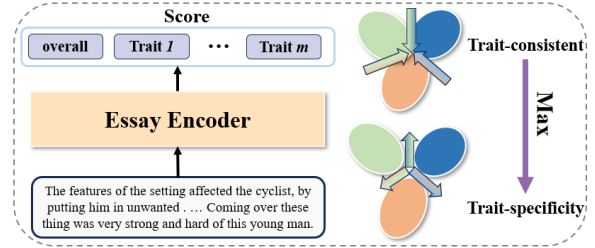


Figure 1: The illustration of multi-trait scoring task and our motivation. Our method aims to enhance the model’s ability to maximally learn trait-specific scoring information within a multi-trait scoring framework.

(Kumar et al., 2022a; Do et al., 2024a) offer greater practical value in real-world educational scenarios.

Earlier work on multi-trait scoring typically utilized the predicted scores of individual traits as supplementary signals to improve the estimation of the holistic essay score (Kumar et al., 2022b). In current multi-trait scoring research, many studies develop joint scoring models that aim to optimize the relationships among different traits, particularly by modeling the dependencies between them (Do et al., 2023, 2024a,b; Li and Ng, 2025; Zhao et al., 2026). This emphasis arises from the observation that, in the trait-scoring task, the scoring trends of different traits often exhibit a degree of consistency (Li and Ng, 2025). For example, both linguistic features and structural organization can reflect content richness in writing assessments.

However, this emphasis on modeling trait inter-dependencies introduces an overlooked challenge: when the learning process overly focuses on scoring dependency across traits, it may obscure the inherent scoring differences among them, ultimately limiting the model’s capacity for multi-dimensional assessment. In other words, if the model over-relies on shared signals, it may under-express features unique to each trait, restricting its capacity to learn trait-specific scoring features.

To alleviate this issue, existing studies have pro-

*Corresponding author.

posed methods to enhance trait-aware scoring representations. For example, (Chu et al., 2024) incorporates Large Language Models (LLMs) to generate trait-specific rationales or label signals, thereby improving the model’s ability to capture trait-relevant scoring information. However, these methods heavily rely on sequential interactions among traits, where the modeling of each trait is conditioned on the outputs of preceding traits. Consequently, they only share knowledge in a fixed, unidirectional order and fail to learn highly expressive trait features (Zhao et al., 2026). Another of method have explored the use of mixture-of-experts architectures to enhance the modeling of trait-specific scoring representations and reduce the interference caused by shared scoring information (Wang and Liu, 2025). However, such attention-based trait representation learning methods are influenced by their optimization objectives, which implicitly introduce scoring consistency signals and hinder the full exploitation of trait-specific features. So, this naturally raises two key questions: 1) How to further enhance the expressiveness of trait-specific features within a unified shared encoder? 2) How to strengthen the focus on trait-specific scoring information while modeling scoring consistency?.

To address this issue, we introduce a Dual-Perspective decoupled Representation learning method for multi-trait Essay Scoring (DPRES), a novel specificity-enhanced multi-trait essay scoring framework. By introducing two decoupled representation learning strategies at the feature representation level and the optimization objective level, our method enables the model to disentangle shared and trait-specific scoring signals without requiring additional trait-specific supervision. Meanwhile, it effectively models inter-trait dependencies while preserving scoring consistency and maintaining sensitivity to scoring differences across traits (Our motivation as illustrated in Figure 1.).

To enhance the model’s ability to capture trait-specific scoring features, we propose a novel residual subspace decomposition strategy, which performs feature-level decoupling by enforcing orthogonality between shared and trait-specific representations. *This method is inspired by the intuition that the score distributions of different traits originate from a common underlying scoring distribution, while exhibiting distributional shifts along different evaluation dimensions. Therefore, enhancing the distinction between the shared signal and these shifted signals can improve the model’s ability to*

perceive trait-specific scoring features. Based on the subspace decomposition learning framework, we further introduce a consistency–specificity decoupled learning strategy. By separating the optimization objectives of scoring-consistency and trait-specificity, this strategy encourages the model to preserve inter-trait dependencies while maximizing its ability to capture trait-specific scoring signals. This decoupled optimization mechanism ensures that the model learns comprehensive and well-balanced trait-scoring representations. The main contributions are summarized as follows:

1) We propose a novel multi-trait essay scoring framework that introduces a dual-perspective decoupled representation learning mechanism to enhance the model’s representation learning capability, thereby significantly improving its ability to model trait-scoring information.

2) We propose a residual subspace decomposition strategy that enforces orthogonality between shared scoring features and trait-specific features, thereby enhancing the model’s ability to perceive initial trait-specific features.

3) We design a consistency–specificity decoupled learning strategy, which decouples the optimization objectives of scoring consistency and trait-specific, enabling the model to preserve trait dependencies while strengthening the learning of trait-specific scoring signals.

4) Extensive experiments on public datasets demonstrate the effectiveness of our method.

2 Related Work

2.1 Automated Essay Scoring

Early studies on Automated Essay Scoring primarily relied on feature engineering or conventional neural architectures, such as CNNs and LSTMs, to predict the overall writing quality of essays (Phandi et al., 2015; Alikaniotis et al., 2016; Mathias and Bhattacharyya, 2018). For example, Alikaniotis et al. (2016) introduced word embeddings and employed BiLSTM networks to capture long-range contextual dependencies in essays, Dong et al. (2017) further proposed a hierarchical representation method with attention mechanisms to model essay structure. With the development of pre-trained language models, Transformer-based approaches have further improved AES performance by leveraging pre-trained contextual representations (Yang et al., 2020; Wang et al., 2022; Xie et al., 2022).

In addition, some studies investigate the cross-

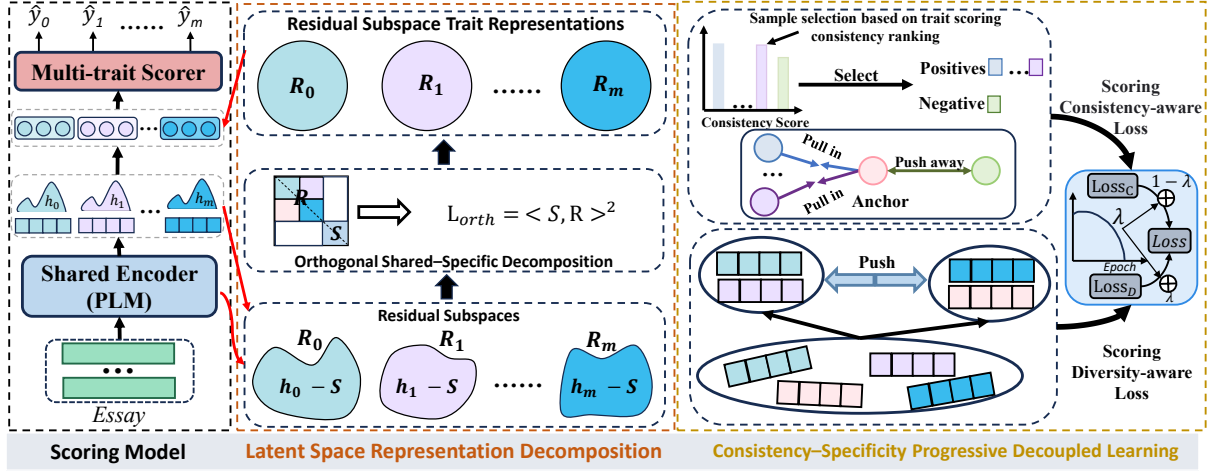


Figure 2: The overall architecture of our method. During multi-trait scoring training, DPRES introduces feature-level decoupled learning through residual subspace decomposition, separating shared scoring features from trait-specific features. The shared embeddings capture common scoring principles, while the residual embeddings preserve trait-specific signals. DPRES further introduces consistency–specificity decoupled learning to separately optimize scoring consistency and trait specificity, preserving inter-trait dependencies while enhancing trait-specific scoring.

prompt essay scoring setting, where models are trained on essays from multiple source prompts and tested on essays written for previously unseen target prompts (Chen and Li, 2023; Jiang et al., 2023; Sun et al., 2025; Wang et al., 2025).

Despite these advances, most existing approaches focus on holistic scoring, which provides limited diagnostic feedback for educational applications. In practice, learners need feedback across multiple writing dimensions to better understand their strengths and weaknesses. This limitation has motivated increasing research on multi-trait essay scoring, which evaluates essays from multiple dimensions and provides more informative feedback.

2.2 Multi-trait Essay Scoring

Multi-trait essay scoring aims to provide feedback across multiple scoring dimensions, enhancing the diagnostic and educational value of AES systems (Ridley et al., 2021; Li and Ng, 2025; Wang and Yu, 2025; Chen et al., 2025). Most existing approaches adopt a unified model with shared essay representations, jointly predicting trait scores and overall scores via multiple linear heads Li and Pan (2025); Wang and Liu (2025); Zhao et al. (2026). Another line of work formulates multi-trait scoring as a sequential generation task: Do et al. (2024b,a) fine-tuned T5-based models to autoregressively generate trait scores, allowing the model to capture correlations among traits during decoding.

Most existing studies focus on modeling inter-trait dependencies to facilitate cross-trait collabo-

rative scoring (Do et al., 2024a; Chu et al., 2024; Li and Ng, 2025; Zhao et al., 2026). However, overemphasizing such dependencies may limit the model’s ability to capture trait-specific scoring information. In contrast, our work takes a decoupling perspective, aiming to disentangle shared scoring consistency from trait-specific representations for more fine-grained multi-trait evaluation.

3 Method

3.1 Overview

The overall architecture of our method is illustrated in Figure 2. In our method, we first construct a unified representation learning framework for multi-trait essay scoring, in which essays are encoded into a shared latent space to capture common scoring representation across different traits. Based on this shared representation, trait-specific scoring heads are introduced to generate trait-level scoring predictions. Second, to enhance the model’s ability to perceive trait-specific scoring signals, we introduce a residual subspace decomposition strategy that explicitly separates shared scoring features from trait-specific representations. Finally, to further balance inter-trait collaboration and the modeling of trait-specific information, we propose a consistency–specificity decoupled learning strategy. Through the joint modeling of decoupling at both the feature- and optimization-level, the proposed DPRES framework effectively captures shared scoring knowledge while enhancing trait-specific repre-

sensation learning in multi-trait essay scoring.

3.2 Multi-trait Scoring Architecture

Our scoring model, denoted as $F(\cdot)$, consists of two components: a shared encoder $f_\theta(\cdot)$ with parameters θ , and a multi-trait scorer $f_\phi(\cdot)$ with parameters ϕ . For the shared encoder, we employ an encoder-based pre-trained language model (PLM), such as BERT (Devlin et al., 2019) or RoBERTa (Liu, 2019), to encode the input essay. Given an essay x , the PLM produces a contextualized representation from the $[CLS]$ token, which serves as the essay-level representation:

$$h_x = f_\theta(x) \quad (1)$$

where h_x denotes the latent representation.

To address the requirement of multi-trait scoring, previous studies (Ridley et al., 2021; Do et al., 2023) have shown that effectively leveraging the mutual information between different traits can significantly improve the performance of trait scoring. Following their work, we construct independent feature projection layers (fully connected layers) for each trait to map the shared essay representation h_x into the initial representation of each trait:

$$a_k = W_k h_x + b_k, \quad k = 1, \dots, m \quad (2)$$

where m denotes the total number of scoring traits, and a_k represents the initial trait-specific representation for the k -th trait.

Subsequently, to capture interaction features across different traits, we introduce a trait interaction network to dynamically extract relevant auxiliary information from other traits when predicting the score of a particular trait. Inspired by the trait attention mechanism (Ridley et al., 2021), our objective is to model the dependencies among traits and enhance each target trait with complementary information from the remaining traits.

Specifically, we first organize the initial representations of all traits into a matrix $A = [a_1, a_2, \dots, a_m]$, where a_k denotes the representation of the k -th trait. Instead of directly aggregating information from other traits using pairwise attention, we employ multiple interaction networks to capture diverse patterns of cross-trait dependencies. Each expert transforms the trait representations while masking the corresponding trait itself, thereby encouraging the extraction of auxiliary information from the remaining traits. Since different traits may benefit from different types of cross-trait

information, a trait interaction weight mechanism is further introduced to dynamically select and aggregate the outputs of the interaction networks for each target trait. In this way, the model can adaptively identify the most relevant interaction patterns and construct an auxiliary representation p_k for the k -th trait. Finally, we concatenate the original trait representation with the aggregated cross-trait information to obtain the fused representation:

$$g_k = [a_k; p_k]. \quad (3)$$

Finally, the representation is fed into a fully connected layer followed by a sigmoid activation function to predict the score of the k -th trait $\hat{y}_k$:

$$\hat{y}_k = \text{Sigmoid}(W_s g_k + b_s), \quad k = 1, \dots, m \quad (4)$$

where W_s and b_s are learnable parameters of the scoring layer. The model ultimately outputs a vector of predicted scores for all traits, denoted as $\hat{y} = [\hat{y}_1, \dots, \hat{y}_m]$.

3.3 Latent Space Representation Decomposition

Separating trait-specific features from a unified $[CLS]$ representation inevitably introduces entanglement among different traits, making it difficult for trait-specific extractors to focus on features relevant to the target trait. Motivated by the intuition that *the scoring distributions of different traits can be viewed as trait-specific shifts from an underlying shared distribution—where the shared distribution captures general writing quality and each trait corresponds to a distinctive distributional shift*, we argue that *explicitly separating shared scoring signals from these trait-specific shifts is essential for accurate trait-level modeling*. To this end, we propose a latent-space representation decomposition strategy. Specifically, we construct residual subspaces to capture trait-specific representations from the shared essay representation, and enforce orthogonality constraints between the shared and trait-specific subspaces to disentangle shared scoring semantics from trait-specific representations.

Residual Subspaces Trait-feature Extracting. After obtaining the trait-specific representations a_k through independent projection layers, we further apply a fully connected layer to project the shared encoder output h_x in order to capture the scoring semantics shared across different traits:

$$z_s = W_s h_x + b_s \quad (5)$$

where z_s denotes the shared essay representation,

which models the common scoring characteristics across different scoring traits.

Based on this shared representation, we construct a residual subspace for the k -th trait by defining the residual embedding as:

$$r_k = a_k - z_s \quad (6)$$

where r_k denotes the residual subspace corresponding to the k -th scoring trait, and a_k represents the initial representation of the k -th trait. According to this definition, the residual embedding preserves the information not captured by the shared representation, thereby encoding trait-specific scoring signals.

Orthogonal Disentanglement of Shared-Representation and Trait-Representation.

Inspired by (Tang et al., 2026), to further learn disentangled shared and trait-specific feature spaces, we introduce a latent-residual orthogonal (LRO) loss. This objective enforces orthogonality between the shared representation z_s and the residual trait embedding matrix $r = [r_1, r_2, \dots, r_m]$. Such a constraint prevents redundant information from being simultaneously encoded in both representations and explicitly separates shared scoring semantics from trait-specific signals.

Formally, the LRO loss is defined as:

$$\mathcal{L}_{LRO} = \sum_{k=1}^m \langle z_s, r \rangle^2 \quad (7)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product operation. The LRO loss penalizes the inner product between the shared representation and all residual embeddings, encouraging their orthogonality and promoting disentangled representation learning. This orthogonality constraint enables a disentangled latent space, where shared representations capture common scoring semantics and residual subspaces encode discriminative trait-specific features.

3.4 Consistency-Specificity Decoupled Learning

Existing studies have shown that scoring dependencies exist across different traits (Do et al., 2023; Li and Ng, 2025; Wang and Liu, 2025). Modeling such dependencies as an optimization objective can improve cross-trait collaboration and enhance prediction consistency. However, explicitly enforcing these dependencies may introduce spurious correlations, which can interfere with trait-specific signals and lead to negative cross-trait interference.

Therefore, building upon the subspace decomposition learning strategy, we further introduce a consistency-specificity decoupled learning strategy. Specifically, we first introduce two complementary learning objectives. *Scoring Consistency-aware Learning* encourages the model to capture shared scoring patterns across different traits, while preserving the inherent correlations among trait scores. *Scoring Diversity-aware Learning* explicitly enhances the discriminability of trait-specific representations, enabling the model to better capture distinctive scoring signals for each trait. Building upon these objectives, we adopt a *progressive decoupling* training strategy: during the early stages of training, the model emphasizes capturing scoring consistency, thereby learning cross-trait scoring collaboration knowledge; as training progresses, the model gradually increases its focus on trait-specific features, improving its sensitivity to unique trait-level signals while maintaining cross-trait collaborative capabilities.

Scoring Consistency-aware Learning. For a given anchor trait representation z_i with score vector s_i , following Do et al. 2023, we first measure its scoring consistency with all other traits based on the Pearson correlation coefficient (PCC) (Benesty et al., 2009). After computing the scoring consistency between each trait and the current anchor trait, we rank them in descending order based on their PCC. The trait with the lowest correlation is selected as the negative sample z_{neg} , and the remaining traits are treated as positive samples $\{z_{\text{pos}}^1 \dots z_{\text{pos}}^{t-2}\}$. This ensures that the model distinguishes the anchor from traits that are less related while reinforcing the cooperative relationships with moderately or highly correlated traits. The contrastive loss for the anchor trait is defined as:

$$\mathcal{L}_{\text{SCL}} = -\log \left(\frac{\sum_{j=1}^{t-2} \exp(\cos(z, z_{\text{pos}}^j) / \tau)}{\sum_{j=1}^{t-2} \exp(\cos(z, z_{\text{pos}}^j) / \tau) + \exp(\cos(z, z_{\text{neg}}) / \tau)} \right) \quad (8)$$

where $\cos(\cdot)$ denotes cosine similarity and τ is a temperature parameter. t means the number of traits. By explicitly ranking traits based on their score correlations and optimizing this contrastive objective, the model learns to emphasize cooperative patterns among traits.

Scoring Diversity-aware Learning. To enhance the model’s ability to learn trait-specific scoring information, we explicitly consider the inherent consistency among traits while promoting trait-specific

distinctions. We introduce a scoring diversity-aware contrastive learning loss that preserves inter-trait consistency and, at the same time, encourages the model to capture diverse, trait-specific features. Specifically, trait representations with high score consistency are separated from other traits, and disentanglement is performed through trait grouping, enabling the model to maximize the learning of distinctive trait characteristics.

Specifically, to filter out the sample pairs with the highest and lowest trait scoring consistency, we first use cosine similarity to quantify the similarity between trait representations. This method quantifies score consistency across different dimensions by enforcing representational coherence, facilitating the disentanglement of traits with substantial scoring disparities in the vector space. For each essay trait representation z being the anchor, the scoring diversity-aware contrastive loss (SDL) is defined as

$$\mathcal{L}_{\text{SDL}} = -\log \left(\frac{\exp(\cos(z, z_{\text{pos}})/\tau)}{\exp(\cos(z, z_{\text{pos}})/\tau) + \exp(\cos(z, z_{\text{neg}})/\tau)} \right) \quad (9)$$

where z represents the current trait representation, z_{pos} denotes the representation most similar to the current trait representation, and z_{neg} denotes the representation least similar to z .

Consistency–Specificity Progressive Decoupled Loss. To better balance scoring consistency and trait-specific learning, we introduce a progressive decoupled optimization strategy. Rather than treating these objectives equally throughout training, the model dynamically adjusts their relative importance to progressively disentangle shared and trait-specific signals. Specifically, the overall training objective is defined as:

$$\mathcal{L}_{\text{CSPD}} = \lambda \mathcal{L}_{\text{SCL}} + (1 - \lambda) \mathcal{L}_{\text{SDL}} \quad (10)$$

where $\mathcal{L}_{\text{SCL}}$ encourages the model to capture trait-consistent scoring patterns across different traits, while $\mathcal{L}_{\text{SDL}}$ promotes the learning of discriminative trait-specific representations.

To progressively decouple these two learning signals, the trade-off parameter λ is dynamically adjusted during training as:

$$\lambda = 1 - \left(\frac{T}{T_{\text{max}}} \right)^2 \quad (11)$$

where T denotes the current training epoch and T_{max} is the maximum number of training epochs. This scheduling strategy allows the model to initially focus on learning scoring consistency and gradually shift its attention toward capturing trait-specific distinctions, thereby enabling a more effective

decoupled representation learning process.

Finally, the overall loss is obtained by combining the Mean Squared Error (MSE) loss with the proposed LRO loss and consistency–specificity progressive decoupled (CSPD) loss,

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha \mathcal{L}_{\text{LRO}} + \beta \mathcal{L}_{\text{CSPD}} \quad (12)$$

where α and β are hyperparameters.

4 Experiments Results

4.1 Experiment Setup

We conducted experiments using the ASAP and ASAP++ datasets (Mathias and Bhattacharyya, 2018), which have been widely used for the MES task. Following previous methods, we adopt the widely used Quadratic Weighted Kappa (QWK) metric for performance evaluation. We use five-fold cross-validation using the same split as previous works (Kumar et al., 2022b), and report the five-fold averaged results. All experiments were conducted on two NVIDIA GeForce RTX 3080 GPUs. In this study, we trained a total of 50 epochs, and the one with the highest average QWK score for all traits in the validation set was selected for testing. We set the max length of the essay is 512 words and the batchsize is 16. In the total loss, we set α as 0.1 and β as 0.01. The drop rate is set to 0.2. For our method, we use RoBERTa-base (Liu, 2019) as the encoder to obtain the multi-dimensional scoring representations. Our model can score all essay traits with a single model. We normalize all trait scores to the range of [0, 1] during training, and the scores are rescaled back to the original score range for evaluation. And we use AdamW (Loshchilov and Hutter, 2017) as our optimizer to train the model, and the initial learning rate is set to $1e - 5$. The temperature for both contrastive losses is set to 0.1. More details about datasets and baselines are presented in Appendix A.

4.2 Main Comparison Results

Table 1 and Table 2 report the main results of the baseline models and our proposed method (DPRES) across prompts and traits, respectively. Overall, our approach consistently outperforms the performance of existing methods on both prompt-level and trait-level. As shown in Table 2, DPRES attains the highest average score (AVG) across traits, demonstrating its robustness across different scoring dimensions. And our method also demonstrates its effectiveness in prompt-level scoring, as

Model	P1	P2	P3	P4	P5	P6	P7	P8	AVG↑ (SD ↓)
MTL-BiLSTM	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665 (-)
T-MES	0.728	0.684	0.702	0.771	0.726	0.754	<u>0.755</u>	0.629	0.719 (-)
GAT-AES	0.722	0.698	0.707	0.771	0.720	0.767	0.744	0.680	0.726 (± 0.009)
EnGNN	<u>0.732</u>	<u>0.704</u>	<u>0.719</u>	0.782	0.734	0.771	<u>0.755</u>	0.636	<u>0.729</u> (-)
ArTS	0.708	0.706	0.704	0.767	0.723	0.776	0.749	0.603	0.717 (± 0.025)
SaMRL	0.717	0.703	0.715	<u>0.773</u>	0.729	<u>0.778</u>	0.745	0.604	0.720 (± 0.016)
RMTS	0.716	<u>0.704</u>	0.723	0.772	<u>0.737</u>	0.769	0.736	0.651	0.726 (± 0.042)
DPRES	0.736	0.719	0.714	<u>0.776</u>	0.740	0.782	0.764	<u>0.668</u>	0.738 (± 0.019)

Table 1: Average QWK scores across all traits for each **prompt**. Five-fold averaged standard deviation is reported (SD). The best results are highlighted in bold and the second-best results are underlined.

Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑ (SD ↓)
MTL-BiLSTM	0.764	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.638 (-)
T-MES	<u>0.774</u>	0.730	0.750	0.702	0.730	0.685	0.686	0.679	0.675	0.693	0.590	0.700 (-)
GAT-AES	0.771	<u>0.742</u>	0.749	0.687	0.726	<u>0.694</u>	0.686	<u>0.709</u>	0.692	0.699	0.649	0.710 (± 0.011)
EnGNN	0.770	0.736	<u>0.759</u>	<u>0.708</u>	0.744	0.691	0.686	0.691	<u>0.699</u>	0.714	<u>0.623</u>	<u>0.711</u> (-)
ArTS	0.737	0.727	0.751	0.702	0.739	0.665	0.679	0.672	0.679	0.735	0.557	0.695 (± 0.018)
SaMRL	0.750	0.732	0.754	0.704	<u>0.740</u>	0.670	0.684	0.681	0.685	0.726	0.558	0.699 (± 0.022)
RMTS	0.755	0.737	0.752	0.713	0.744	0.682	<u>0.690</u>	0.705	0.694	0.702	0.612	0.708 (± 0.043)
DPRES	0.777	0.755	0.763	<u>0.705</u>	<u>0.740</u>	0.705	0.711	0.715	0.701	<u>0.728</u>	0.649	0.722 (± 0.020)

Table 2: Average QWK scores across all prompts for each **trait**. WC: Word Choice, Org: Organization, SF: Sentence Fluency, Conv: Conventions, PA: Prompt Adherence, Nar: Narrativity, Lang: Language.

presented in Table 1.

In particular, our method surpasses the strongest baseline model EnGNN by 1.1% on average from the perspective of trait-level scoring, demonstrating the effectiveness of the proposed framework in multi-trait essay scoring. Statistical significance tests further validate these results, revealing consistent improvements of 0.009 and 0.011 at the prompt and trait levels, both of which are statistically significant ($p < 0.05$). Further compared with methods that focus on modeling trait dependencies, including GAT-AES, ArTS, and SaMRL, our method achieves significant advantages both at the prompt-level and the trait-level. Compared to trait-aware baselines such as T-MES and RMTS, our approach also demonstrates significant advantages. Notably, against RMTS, which leverages LLMs to generate supplementary trait information, our decoupled learning framework achieves superior performance, highlighting the efficacy of our architectural design in capturing trait-specific representations.

4.3 Ablation studies

We conduct ablation studies to evaluate the contribution of each component, with results summarized in Tables 3 and 4. (1) w/o LRO: Removing the LRO constraint reduces performance from 0.738 to 0.726, indicating that enforced orthogonality effectively mitigates trait-representation entanglement

and enhances trait-specific discriminability. (2) w/o CSPD: Ablating the CSPD module yields a drop to 0.729, demonstrating that progressively balancing shared and trait-specific representations during training is crucial for robust optimization. (3) w/o LRO & CSPD: Joint removal causes the most substantial degradation (0.738 $\rightarrow$ 0.714), underscoring the complementary nature of both mechanisms. In addition, at the trait-level, our method consistently improves performance across most traits, with notable gains on *Style*, *Voice*, and *Sentence Fluency*, highlighting the importance of modeling trait-specific representations for fine-grained scoring. Overall, the complete model achieves the best performance, validating the effectiveness of jointly addressing trait-representation entanglement and consistency-specificity balance.

4.4 More Analysis

Effect of Optimization Order on Consistency-Specificity Decoupled Learning. We investigate the effect of the optimization order in consistency-specificity decoupled learning across different prompts by comparing two sequences: training scoring consistency first followed by prompt-specific learning (Consistency $\rightarrow$ Specificity) and the reverse order (Specificity $\rightarrow$ Consistency). As shown in Figure 3, training consistency first consistently outperforms the reverse order. Specif-

Model	P1	P2	P3	P4	P5	P6	P7	P8	AVG↑ (SD ↓)
DPRES	0.736	0.719	0.714	0.776	0.740	0.782	0.764	0.668	0.738 (± 0.019)
w/o LRO	0.731	0.700	0.708	0.775	0.727	0.770	0.754	0.642	0.726 (± 0.026)
w/o CSPD	0.727	0.718	0.714	0.776	0.724	0.768	0.759	0.651	0.729 (± 0.022)
w/o LRO & CSPD	0.721	0.673	0.708	0.768	0.717	0.762	0.748	0.618	0.714 (± 0.024)

Table 3: Ablation results in average QWK of all traits on each prompt.

Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑ (SD ↓)
DPRES	0.777	0.755	0.763	0.705	0.740	0.705	0.711	0.715	0.701	0.728	0.649	0.722 (± 0.020)
w/o LRO	0.767	0.747	0.754	0.690	0.728	0.692	0.677	0.679	0.670	0.713	0.596	0.702 (± 0.030)
w/o CSPD	0.773	0.745	0.758	0.693	0.727	0.701	0.699	0.698	0.687	0.715	0.620	0.710 (± 0.024)
w/o LRO & CSPD	0.763	0.730	0.755	0.679	0.722	0.675	0.679	0.662	0.652	0.675	0.575	0.688 (± 0.028)

Table 4: Average QWK scores across all prompts for each trait.

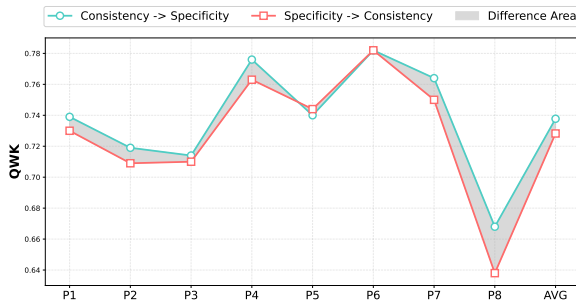


Figure 3: Effect of the optimization order on consistency–specificity decoupled learning. Consistency $\rightarrow$ Specificity indicates that, under the consistency–specificity progressive decoupled (CSPD) loss, scoring consistency is trained first, followed by enhanced learning of trait-specific representations, Specificity $\rightarrow$ Consistency reverses the order.

ically, the average performance across prompts is higher for Consistency $\rightarrow$ Specificity, and the lowest-performing prompt also shows a clear improvement. Most prompts exhibit better scores under Consistency $\rightarrow$ Specificity, with the largest improvement observed in prompt 8. These results further validate the motivation behind our method: overemphasizing scoring consistency can impede the model’s ability to perceive prompt-specific scoring signals, whereas the proposed progressive decoupled objective effectively facilitates joint modeling of scoring consistency and prompt-specific information.

Robustness Analysis of the Optimization-Level Dynamic Progressive Decoupling Mechanism. To evaluate the robustness of the proposed optimization-level dynamic progressive decoupling mechanism, we conducted an ablation study on different progressive scheduler functions. Five scheduler functions were considered, and the results are

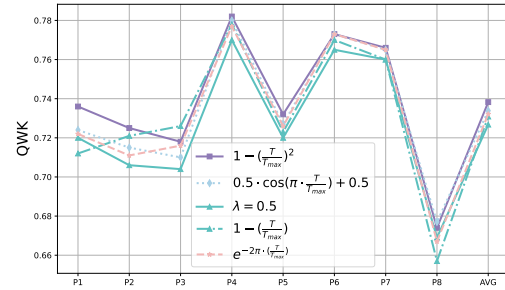


Figure 4: Effect of different scheduler functions on the proposed consistency–specificity decoupled learning.

shown in Figure 4. The top panel illustrates the weight curves of each scheduler, while the bottom panel presents the corresponding overall scoring performance. The results indicate that variations in scheduler design have minimal impact on model performance. In particular, the differences in final averaged scores are negligible, demonstrating the robustness of our method to the choice of progressive scheduler. Furthermore, the model performs worst under a fixed schedule ($\lambda = 0.5$), which further indicates that, compared to static weight allocation, the proposed progressive decoupling strategy is more effective in improving multi-trait scoring performance.

t-SNE Visualization of Shared and Trait-specific Representations. To evaluate the effectiveness of the latent residual orthogonalization (LRO) loss, we perform t-SNE visualization on P1-P4 prompts in ASAP/ASAP++. The spatial distributions of shared and trait-specific representations are shown in Figure 5. Without LRO (top panels), severe trait-representation entanglement is observed: shared and trait-specific representa-

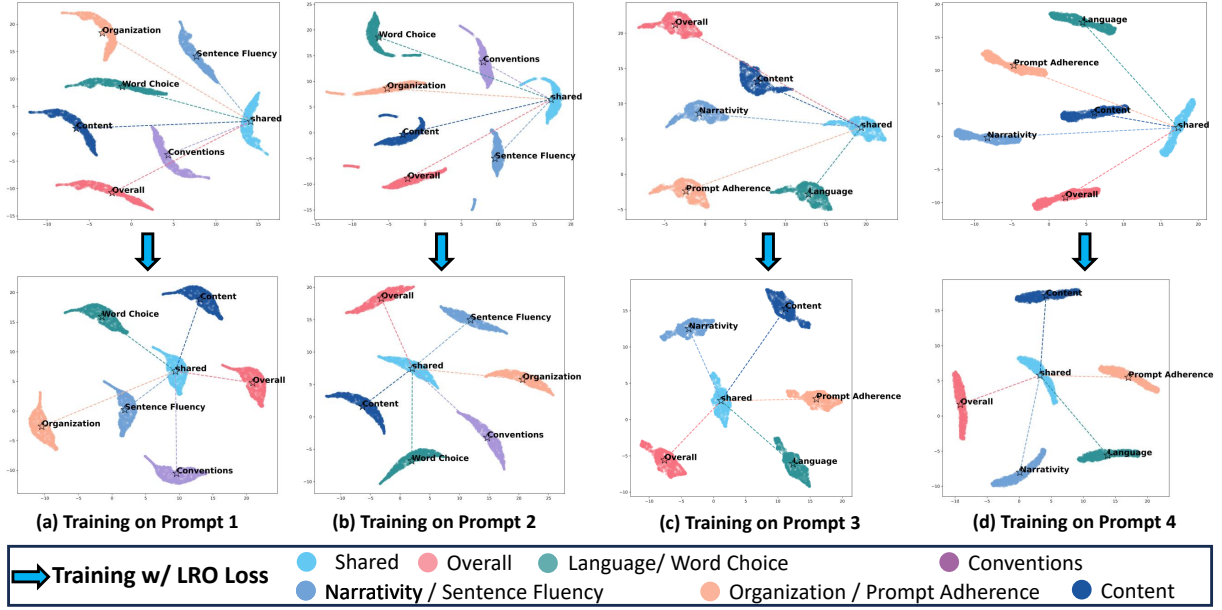


Figure 5: t-SNE visualization of trait-specific scoring representations and shared essay representations across P1-P4 prompts. After training with the Latent-Residual Orthogonal (LRO) loss, the latent space exhibits clearer separation between shared and trait-specific representations.

tions overlap significantly, and clusters of different traits are poorly separated, indicating strong cross-trait interference. With LRO (bottom panels), representations become clearly disentangled. Shared representations form a central core, while trait-specific features are distributed in distinct directions with compact and well-separated clusters. These results demonstrate that LRO effectively disentangles shared and trait-specific representations, thereby improving the model’s ability to capture discriminative trait-specific scoring signals.

5 Conclusion

In this work, we propose DPRES, a dual-perspective decoupled representation learning framework for multi-trait essay scoring. DPRES addresses the inherent tension between cross-trait consistency and trait-specific discrimination by disentangling shared and trait-specific representations at both the feature and optimization levels. It employs a residual subspace decomposition strategy to separate shared scoring semantics from trait-specific features, and a consistency–specificity decoupled learning strategy to balance inter-trait correlations with trait-level discrimination. Extensive experiments on public datasets demonstrate that DPRES consistently outperforms strong baselines across multiple prompts and trait dimensions.

Limitations

Despite our best efforts, our study may still have several limitations. First, since our method achieves maximization of inter-trait scoring differences solely from a contrastive learning perspective, future exploration of more diverse disentanglement strategies may further improve trait-level scoring performance. Second, due to dataset limitations, our method has only been evaluated on English datasets; testing it in multilingual settings in the future would further validate its generalization capability. Additionally, as suggested by existing work, latent multi-hop dependencies exist among different traits (Li and Ng, 2025). Therefore, further uncovering these deeper inter-trait dependencies while simultaneously enhancing the model’s focus on trait-specific scoring information may lead to more effective cross-trait scoring inference.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 62477002). During the preparation of this work, the authors only used generative-AI tools for grammar-checking purposes.

References

- Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. *ArXiv*, abs/1606.04289.
- Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.
- Po-Kai Chen, Bo-Wei Tsai, Shao Kuan Wei, Chien-Yao Wang, Jia-Ching Wang, and Yi-Ting Huang. 2025. Mixture of ordered scoring experts for cross-prompt essay trait scoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18071–18084.
- Yuan Chen and Xia Li. 2023. **PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.
- SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. 2024. Rationale behind essay scores: Enhancing s-llm’s multi-trait essay scoring with rationale generated by llms. *arXiv preprint arXiv:2410.14202*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Heejin Do, Yunsu Kim, and Gary Lee. 2024a. **Autoregressive score generation for multi-trait essay scoring**. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian’s, Malta. Association for Computational Linguistics.
- Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. **Prompt- and trait relation-aware cross-prompt essay trait scoring**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.
- Heejin Do, Sangwon Ryu, and Gary Lee. 2024b. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438.
- Fei Dong, Yue Zhang, and Jie Yang. 2017. **Attention-based recurrent convolutional neural network for automatic essay scoring**. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.
- Ahmed M Elmassry, Nazar Zaki, Negmeldin Alsheikh, and Mohammed Mediani. 2025. A systematic review of pretrained models in automated essay scoring. *IEEE Access*, 13:121902–121917.
- Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470.
- Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022a. **Many hands make light work: Using essay traits to automatically score essays**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495, Seattle, United States. Association for Computational Linguistics.
- Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022b. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495.
- Shengjie Li and Vincent Ng. 2025. **Graph-based multi-trait essay scoring**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33313–33339, Suzhou, China. Association for Computational Linguistics.
- Xia Li and Wenjing Pan. 2025. Kaes: Multi-aspect shared knowledge finding and aligning for cross-prompt automated scoring of essay traits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24476–24484.
- Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2017. **Decoupled weight decay regularization**. In *International Conference on Learning Representations*.
- Sandeep Mathias and Pushpak Bhattacharyya. 2018. **ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- P. Phandi, Kma Chai, and H. T. Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of AAAI Conference on Artificial Intelligence*.
- Jingbo Sun, Tianbao Song, Weiming Peng, and Jihua Song. 2025. A survey of automated essay scoring: Challenges, advances, and future. *Neurocomputing*, 650:130916.
- Kaveh Taghipour and Hwee Tou Ng. 2016. [A neural approach to automated essay scoring](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.
- Xiaole Tang, Xiaoyi He, Jiayi Xu, Xiang Gu, and Jian Sun. 2026. Learning continuous wasserstein barycenter space for generalized all-in-one image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jiong Wang and Jie Liu. 2025. [T-MES: Trait-aware mix-of-experts representation learning for multi-trait essay scoring](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1224–1236, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jiong Wang and Shengquan Yu. 2025. Improving prompt generalization for cross-prompt essay trait scoring from the scoring-invariance perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2633–2646.
- Jiong Wang, Qing Zhang, Jie Liu, Xiaoyi Wang, Mingying Xu, Liguang Yang, and Jianshe Zhou. 2025. Making meta-learning solve cross-prompt automatic essay scoring. *Expert Systems with Applications*, 272:126710.
- Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425.
- Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. [Automated essay scoring via pairwise contrastive regression](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.
- Shiman Zhao, Siyuan Liu, and Zhiqi Shen. 2026. Multi-knowledge enhanced graph neural network for multi-trait essay scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28733–28741.

Prompt	No. of Essays	Traits	Score Range	
			Overall	Trait
1	1,783	Content, Organization, Word Choice, Sentence Fluency, Conventions	2-12	1-6
2	1,800	Content, Organization, WC, Sentence Fluency, Conventions	1-6	1-6
3	1,726	Content, Prompt Adherence, Language, Narrativity	0-3	0-3
4	1,772	Content, Prompt Adherence, Language, Narrativity	0-3	0-3
5	1,805	Content, Prompt Adherence, Language, Narrativity	0-4	0-4
6	1,800	Content, Prompt Adherence, Language, Narrativity	0-4	0-4
7	1,569	Content, Organization, Conventions, Style	0-30	0-6
8	723	Content, Organization, Word Choice, Sentence Fluency, Conventions, Voice	0-60	2-12

Table 5: The detailed statistics of the ASAP/ASAP++ datasets. The prompt is an instruction that defines the writing theme.

A More Experiment Setup

A.1 Dataset

The ASAP dataset contains eight prompts with English essays. Only prompts 7 and 8 have additional trait scores. The ASAP++ dataset, which includes the same essays, extends the ASAP dataset by providing multi-trait scores for prompts 1–6 (as shown in Table 5). Consequently, trait scores for prompts 1–6 are taken from ASAP++, while the scores for prompts 7 and 8, along with overall scores, are sourced from ASAP.

A.2 Baseline Methods

We compare our method with representative discriminative and generative approaches.

Discriminative-based approaches: *MTL-BiLSTM* (Kumar et al., 2022b) is a multi-task scoring model, where the predicted trait scores support the overall scoring process. *T-MES* (Wang and Liu, 2025) is the multi-trait scoring method based on pre-trained language models. *GAT-AES* (Li and Ng, 2025) is a graph-based multi-trait essay scoring method. *EnGNN*¹ (Zhao et al., 2026) is also a graph-based model for multi-trait essay scoring.

Generative-based approaches: *ArTS* (Do et al., 2024a) is an autoregressive multi-trait score generation model based on encoder-decoder (T5) (Raffel et al., 2020). *SamRL* (Do et al., 2024b) further utilizes reinforcement learning on base of ArTS method to facilitate trait score generation. *RMTS* (Chu et al., 2024) uses LLMs to extract rationales as additional contextual input for essays, and applies the T5 model for multi-trait scoring generation.

Table 6: Comparison of DPRES performance using different pre-trained language models.

Encoder	Avg. QWK on prompts	Avg. QWK on traits
BERT	0.730	0.714
RoBERTa	0.738	0.722
XLNet	0.736	0.720

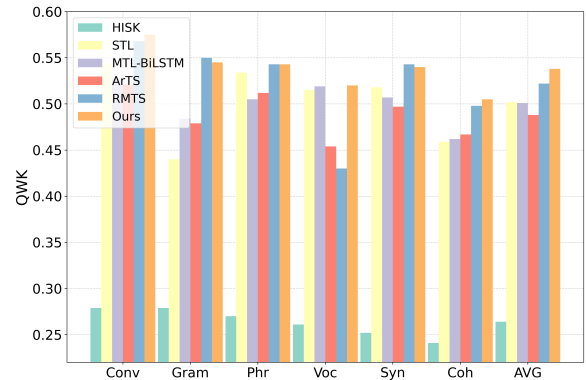


Figure 6: Average QWK scores across all prompts for each trait on the Feedback Prize dataset.

B Additional experimental results

Effect of different pre-trained models. To investigate the effect of encoder choice on trait-specific scoring, we conducted comparative experiments using different encoder-based pre-trained language models, including BERT, RoBERTa, and XLNet. The results are summarized in Table 6. As shown, different pre-trained models exhibit variations in performance, reflecting differences in their encoding capabilities. Importantly, our proposed DPRES framework consistently improves trait-specific scoring performance across all encoders, demonstrating its robustness to the choice of pre-trained language model.

¹As the original paper does not provide a specific model name, we adopt this name for clarity.

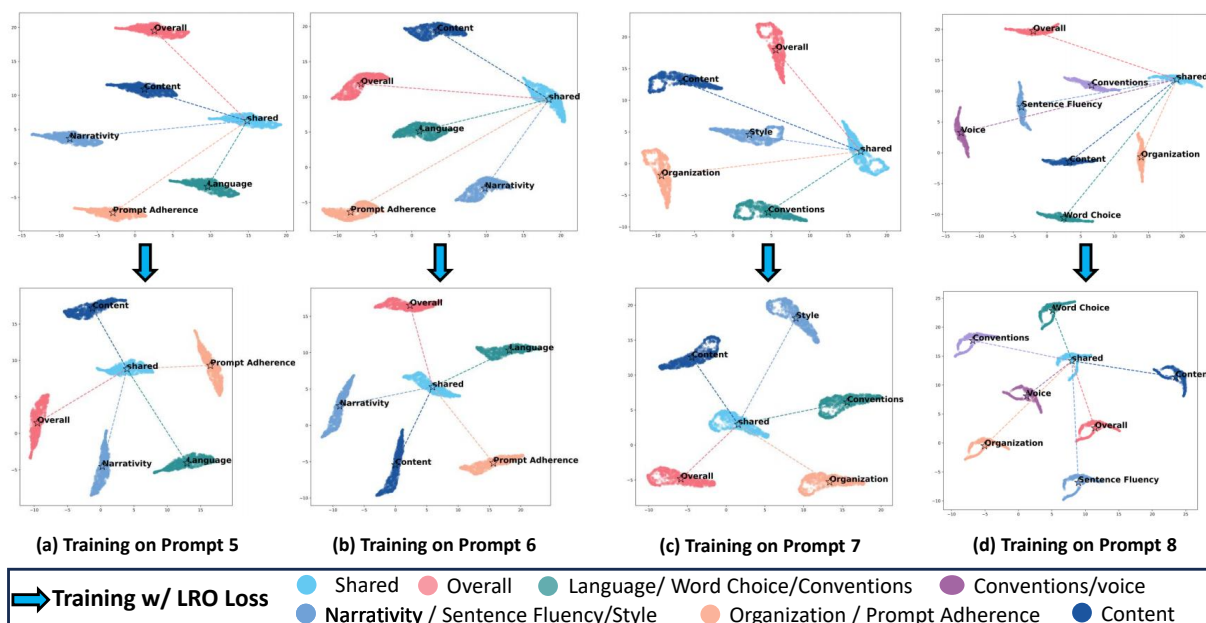


Figure 7: t-SNE visualization of trait-specific scoring representations and shared essay representations across P5-P8 prompts.

Performance with Feedback Prize Dataset. To further demonstrate the advantage of our method in multi-trait scoring, we conducted additional experiments on the Feedback Prize dataset², as presented in Figure 6. The results demonstrate that our method consistently achieves the best average performance, indicating its broader applicability.

Visualization on other prompts. Figure 7 further presents the visualization results on prompts P5–P8. As observed from the figure, the results are consistent with those shown in Figure 5, demonstrating that our method exhibits similar patterns across the remaining prompts as well.

²<https://www.kaggle.com/competitions/feedback-prizeenglish-language-learning/data>