

Article

MTL-Frame: An End-to-End Multi-Task Learning Framework for Student Profiling and Composite-Score Estimation

Guifen Jiang , Muhua Tan  and Zhaohui Yuan * 

School of Foreign Language, East China Jiaotong University, 808 East Shuanggang Rd., Nanchang 330013, China; jiang.guifen@hotmail.com (G.J.); 2024218083500003@ecjtu.edu.cn (M.T.)

* Correspondence: yuanzh@ecjtu.edu.cn

Abstract

Composite-score estimation (or assessment auditing) and student profiling are two fundamental tasks in Educational Data Mining (EDM). However, existing studies often treat them as separate problems, typically adopting a sequential pipeline in which clustering is first performed and then used for downstream estimation. Such a fragmented paradigm limits the interaction between latent student-group structures and supervised outcome signals, particularly in low-data educational scenarios. To address this limitation, this study proposes “MTL-Frame”, an end-to-end multi-task learning framework that jointly optimizes student profiling and composite-score estimation. MTL-Frame integrates prototype-based contrastive clustering with context-aware regression to directly inject student profile priors into grade estimation. A dispersion regularization and dynamic loss weighting ensure training stability. Experiments conducted on a real-world blended English-course dataset involving 429 university students show that MTL-Frame outperforms representative single-task regression baselines, including XGBoost, Random Forest, SVR, and LSTM, achieving an RMSE of 1.7532, an MAE of 0.8168, and $R^2 = 0.9812$. While this high R^2 partly reflects the model’s ability to learn the deterministic scoring aggregation, the performance remains strong even when the final exam score is excluded from the inputs ($R^2 = 0.9681$), confirming genuine estimation capability. Compared to the strongest single-task baseline (XGBoost), MTL-Frame reduces RMSE by 38.2%. The model also obtains a Silhouette Score of 0.3562, indicating its ability to generate meaningful student profiles while maintaining strong estimation accuracy. These results demonstrate that integrating unsupervised profiling priors into supervised estimation can improve model robustness and provide actionable insights for differentiated instructional intervention.

Keywords: educational data mining; learning analytics; student profiling; composite-score estimation; multi-task learning



Academic Editors: Solange Denervaud and Anaëlle Camarda

Received: 23 May 2026

Revised: 23 June 2026

Accepted: 24 June 2026

Published: 27 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Transformation of education, the blended teaching model—integrating online and offline environments—has become the new normal in higher education [1]. This paradigm shift has generated voluminous, multi-source, and heterogeneous educational data, offering unprecedented opportunities for Educational Data Mining (EDM) and learning analytics [2]. A core objective of EDM is to extract high-value patterns from massive behavioral and assessment data to support precise pedagogical decision-making and personalized learning interventions [3]. Consequently, student profiling and composite-score estimation (or

assessment auditing), as two foundational tasks within EDM, have garnered extensive academic attention [4].

However, traditional research frequently adheres to a single-task-driven paradigm or employs a phased approach that strictly separates clustering from subsequent prediction [5,6]. Such a fragmented processing framework neglects the inherent logical correlations within the learning process. In reality, precise student profiling reveals the internal structure of learning patterns, providing prior guidance for Composite-Score Estimation; conversely, accurate prediction signals reflect the external manifestation of learning outcomes, which can inversely validate and refine the clustering boundaries of population profiles [7]. Recently, Multi-Task Learning (MTL) has demonstrated remarkable advantages across various artificial intelligence domains by facilitating inter-task knowledge transfer through shared underlying representations [8]. Integrating unsupervised profile clustering and supervised performance regression into a unified MTL framework holds the potential to overcome the performance bottlenecks of conventional serial paradigms and exploit synergistic gains between tasks [9]. This not only possesses significant theoretical value for exploring new paradigms in educational AI but also provides essential technical scaffolding for constructing trustworthy, intelligent personalized learning support systems [10].

Despite substantial progress in EDM technologies, research focusing on the collaborative optimization of profiling and prediction continues to face several critical challenges:

1. **Lack of Task Synergy and Joint Optimization Mechanisms.** Most existing methods adopt serial processing strategies, treating profile construction and Composite-Score Estimation as mutually independent stages [11]. This architecture prevents the model from learning shared representations that are discriminative for both tasks under a unified optimization objective [12]. Consequently, clustering results fail to provide effective guidance for downstream regression tasks, and the supervisory signals from regression cannot efficiently propagate back to the clustering space to refine population distributions.

2. **Insufficient Deep Fusion of multidimensional formative and summative assessment data.** Data in authentic teaching scenarios exhibits high heterogeneity, encompassing structured grades, semi-structured assignments, and unstructured behavioral logs [13]. Current research predominantly relies on singular features or simplistic vector concatenation, lacking adaptive fusion mechanisms capable of capturing the intrinsic correlations among multimodal educational data [14]. This restricts the model's capacity to comprehensively portray the deep cognitive states of learners.

3. **Weak Interpretability of Model Decision Processes.** Educational applications demand exceptionally high levels of model transparency and logical traceability [15]. Existing deep learning models often struggle to elucidate the allocation logic of shared representations across different tasks, nor can they clearly explain the interaction mechanisms between unsupervised clustering and supervised prediction. This opacity severely restricts their credible deployment and application in real-world educational environments [16,17].

To address the aforementioned challenges, this paper proposes an end-to-end collaborative optimization framework for student profiling and Composite-Score Estimation based on multi-task learning. Unlike conventional deep clustering frameworks primarily developed for image or text data, MTL-Frame is purpose-built for small-sample educational tabular data, incorporating prototype-based clustering and targeted regularization to mitigate overfitting while maintaining high interpretability. Specific research questions guiding this study are: (1) Can the joint optimization of unsupervised clustering and supervised regression yield higher prediction accuracy than single-task models on small-sample educational data? (2) Does the prototype-based clustering branch produce meaningful and interpretable student profiles? (3) How do the core components—clustering branch,

prototype regularization, and dynamic loss weighting—contribute to overall performance and training stability? The primary contributions of this work are summarized as follows:

1. We propose a novel multi-task joint optimization architecture that integrates unsupervised student population clustering and supervised composite-score estimation into a unified end-to-end learning pipeline. The model learns to infer the institution's implicit grade aggregation formula from fine-grained formative indicators, without being provided the weighting scheme. Through the collaborative design of a shared encoder and task-specific decoders, it simultaneously captures deep data structures and explicit supervisory signals, achieving complementary advantages and deep synergy between tasks.
2. A prototype-based clustering and feature-enhancement mechanism. We design a contrastive clustering branch with learnable prototypes to model latent student-group structures. The learned cluster context is further incorporated into the regression branch, enabling profile-aware composite-score estimation.
3. A regularized optimization strategy for stable multi-task training. To reduce prototype collapse and task interference, we introduce prototype dispersion regularization and a dynamic clustering-loss weighting strategy, improving the stability of representation learning under small-sample educational data conditions.
4. Empirical validation on a real-world blended-learning dataset. We evaluate the proposed framework on a university English course dataset containing 429 students and compare it with representative regression and clustering baselines. The results demonstrate the effectiveness of the proposed joint-learning paradigm.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 elaborates on the proposed methodology and technical details; Section 4 introduces the dataset and experimental design; Section 5 discusses the experimental findings; Section 6 provides an in-depth discussion of the results, limitations, and practical implications; and finally, Section 7 concludes the paper and outlines future research directions.

2. Related Work

This section briefly reviews the research progress and existing challenges in educational data mining, student profiling, and the application of multi-task learning within the educational domain.

2.1. Educational Data Mining

Educational Data Mining aims to extract valuable insights from multidimensional formative and summative assessment data to optimize instructional decisions and learning interventions. Early research predominantly relied on traditional machine learning methods, such as linear regression, decision trees, and Support Vector Machines (SVMs), for composite-score estimation. Recently, with the widespread adoption of deep learning, neural network-based predictive models have significantly improved accuracy. For instance, ref. [18] explored novel features influencing undergraduate academic performance through systematic feature engineering, while [19] demonstrated the effectiveness of extracting behavioral logs from online learning management systems for prediction. To capture complex inter-student correlations, advanced models like Graph Neural Networks (GNNs) have been introduced to enhance prediction performance by modeling structural relationships among students or knowledge concepts [20].

Despite continuous improvements in predictive accuracy, most existing academic prediction research follows a single-task paradigm, treating grade prediction as an isolated problem. A systematic review by [21] highlighted that a majority of methods fail to fully utilize the latent structural information present within student populations. Although at-

tempts have been made to integrate unsupervised and supervised techniques—such as conducting clustering first and then utilizing cluster labels as feature inputs for predictive models [22]—this two-stage pipeline fragments the intrinsic connection between tasks. Consequently, the model struggles to learn shared representations with strong discriminative power for downstream tasks in an end-to-end manner. Therefore, exploring a unified framework capable of synergistically optimizing student segmentation and Composite-Score Estimation is critical to overcoming the limitations of current single-task paradigms.

2.2. Student Profiling

Student profiling aims to identify student groups with similar learning characteristics or behavioral patterns through unsupervised or semi-supervised methods, serving as the foundation for differentiated instruction and personalized support. Traditional approaches predominantly rely on classical algorithms like K-Means and hierarchical clustering, which remain widely utilized in educational data analysis due to their simplicity and efficacy [23]. As educational scenarios grow in complexity, researchers have begun exploring more refined profiling techniques. For example, ref. [24] proposed a machine learning-based approach to construct population profiles and predict their classifications, whereas [25] designed an interpretable neural network clustering pipeline that generates personalized feature subsets for different students via adaptive feature selection prior to clustering.

In online learning environments, researchers frequently combine clustering with predictive models. Ref. [26] performed fine-grained classification of online learning records and constructed predictive models to assess learning outcomes, while [27] utilized anomaly detection and explainable AI techniques to discover unconventional learning patterns. However, existing research exhibits two major limitations: First, most profiling efforts conclude at descriptive analysis and fail to effectively integrate the mined population information into downstream predictive tasks, thereby limiting the decision-support value of the profiles. Second, current methods predominantly rely on single data sources (e.g., exclusively grades or behavioral logs), lacking deep fusion and unified representation learning of multimodal educational data, which restricts the comprehensiveness of the generated profiles.

2.3. Multi-Task Learning in Education

Multi-Task Learning aims to enhance model generalization and data utilization efficiency by sharing representations and jointly optimizing multiple related tasks—a paradigm that has achieved notable success in fields like computer vision. Recently, its application has begun to expand into the realm of EDM. For instance, ref. [11] proposed utilizing a multi-task Long Short-Term Memory (LSTM) network to simultaneously process holistic student Composite-Score Estimation and performance level classification tasks. In the specific context of Knowledge Tracing (KT), ref. [28] designed a multi-task model to jointly learn representation vectors for skill difficulty and problem difficulty. More specialized explorations include introducing response speed as an auxiliary task in KT [28], or refining student assessments by combining KT with option tracing [29].

Nevertheless, the application of MTL frameworks for the collaborative optimization of student profiling and Composite-Score Estimation remains in its infancy and faces distinct challenges. First, existing research largely focuses on the joint learning of homogeneous tasks (e.g., multiple classification or regression tasks), with rare explorations into incorporating unsupervised clustering tasks and supervised regression/classification tasks within the same end-to-end optimization framework. Second, network architecture designs often lack specific consideration for the unique characteristics of educational data, such as temporal dynamics, sparsity, and multi-scale nature; most works directly adopt generic multi-task

sharing architectures. Finally, model interpretability remains generally weak. In MTL, how shared representations influence the decision-making processes of different tasks is often opaque—a critical deficiency for educational applications that demand high credibility and interpretability.

3. Methodology

To achieve the collaborative optimization of student profiling and Composite-Score Estimation, we propose an end-to-end unified framework based on multi-task learning. Through a shared representation learning mechanism, this framework integrates the unsupervised clustering task and the supervised regression task into a single training pipeline, enabling the two tasks to mutually reinforce and jointly optimize each other [30]. This section first formally defines the research problem, followed by a detailed elucidation of the model architecture, loss function design, and training optimization strategy.

3.1. Model Architecture

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ containing n students, where $x_i \in \mathbb{R}^d$ denotes the d -dimensional feature vector of the i -th student, and $y_i \in \mathbb{R}$ represents the corresponding final overall grade. This study aims to simultaneously address two critical tasks: student profiling (clustering task) and Composite-Score Estimation (regression task). Specifically, the clustering task partitions students into k meaningful groups in an unsupervised manner, with each group representing specific learning patterns or behavioral traits [31]. The regression task accurately predicts the overall grade y_i based on the student's features x_i [32]. It is important to clarify the relationship between the input features and the target variable. The final overall grade y_i is derived from the same assessment activities represented by the 10 input features, but through a complex, multi-level weighted aggregation scheme that is not provided to the model. The model must learn this aggregation logic from the data. The input consists of 10 fine-grained, atomized metrics, while the target is a single, aggregated score. The model's task is to discover how these individual indicators combine to produce the final grade, a non-trivial learning problem that involves inferring the implicit weighting structure of the course's assessment design. Moreover, the final examination score is only one of the 10 input dimensions; the regression branch learns from all features jointly, and the clustering branch provides additional population-level priors that enhance the estimation. This design does not constitute data leakage: the model is not given the formula for computing the final grade, and the final examination score contributes only a portion of the total information used for estimation.

The overall architecture of MTL-Frame is illustrated in Figure 1. The framework primarily consists of three core components: a Shared Feature Extractor for capturing generalized patterns, a Clustering Branch for mining population structures, and a Regression Branch for outputting the final predictions. The computational workflow and design motivations for each component are detailed below.

3.1.1. Shared Feature Extractor

Given the student input features $x \in \mathbb{R}^d$, which encompass multidimensional formative and summative evaluation metrics, the shared feature extractor aims to learn a high-dimensional abstract representation $h \in \mathbb{R}^{d_h}$ that exhibits strong discriminative power for both downstream unsupervised population segmentation and supervised Composite-Score Estimation. Formally, this mapping process f_θ is defined as:

$$h = f_\theta(x) = \text{ReLU}(W_2 \times \text{ReLU}(W_1 x + b_1) + b_2) \quad (1)$$

where $\theta = \{W_1, W_2, b_1, b_2\}$ represents the learnable weight and bias parameters in the encoder, and d_h is the hidden layer dimensionality. To mitigate overfitting phenomena typically observed when applying deep networks to small-scale educational tabular data, we introduce Batch Normalization and Dropout regularization following the fully connected layers. Unlike traditional single-task models, this hard parameter sharing mechanism forces the network to optimize under the joint propulsion of dual tasks. This compels the encoder to extract generalized global learning behavioral features, effectively preventing the mechanical memorization of localized noise [33].

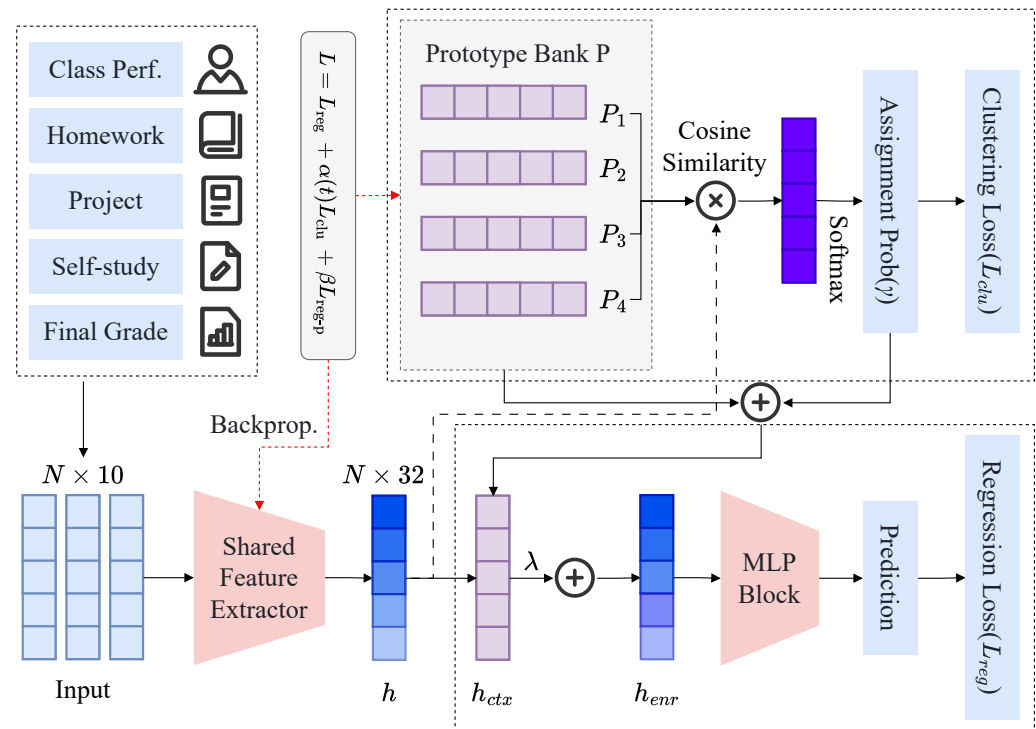


Figure 1. The overall architecture and optimization strategy of MTL-Frame.

3.1.2. Prototype-Based Clustering Branch

To fully excavate the latent semantic structure of the student population, we introduce a contrastive learning-based clustering branch. Specifically, we define a set of learnable cluster prototypes $P = \{p_j\}_{j=1}^k \in \mathbb{R}^{k \times d_h}$ within the latent space, where each prototype p_j characterizes a specific population learning pattern [34]. Given a shared representation h_i , its cosine similarity with each cluster prototype is computed as:

$$s_{ij} = \frac{h_i \times p_j}{\|h_i\| \|p_j\|}, \quad j = 1, \dots, k \quad (2)$$

Subsequently, the similarity distances are mapped into soft cluster assignment probabilities γ_{ij} via a Softmax function:

$$\gamma_{ij} = \frac{\exp(s_{ij}/\tau)}{\sum_{j'=1}^k \exp(s_{ij'}/\tau)} \quad (3)$$

where τ is the temperature parameter controlling the smoothness of the distribution. The final hard cluster label is determined by the maximum probability: $c_i = \arg \max_j \gamma_{ij}$. Unlike traditional cluster-then-predict pipeline methods, the prototype bank P in this branch is updated end-to-end concurrently with network backpropagation. This implies that the student profiles generated by the model reflect not only the geomet-

ric distribution of the feature space but are also strongly corrected by explicit downstream academic outcome signals, thereby endowing the clustering results with profound pedagogical significance [35].

3.1.3. Context-Aware Regression Branch

The regression branch aims to accomplish precise prediction of individual final grades based on the deep fusion of features. We first utilize the cluster assignment probabilities to perform a weighted aggregation of the prototype bank, constructing a cluster context representation h_i^{ctx} specific to that sample:

$$h_i^{\text{ctx}} = \sum_{j=1}^k \gamma_{ij} p_j \quad (4)$$

Subsequently, this population context information is fused with the original individual shared representation via a residual connection to generate an enriched feature representation h_i^{enr} :

$$h_i^{\text{enr}} = h_i + \lambda h_i^{\text{ctx}} \quad (5)$$

where λ is the fusion coefficient. The core motivation behind this design is that an individual student's academic performance is deeply embedded within the common patterns of their respective group. Injecting h_i^{ctx} , which contains global priors, into the individual feature h_i fundamentally leverages population profiles to recalibrate and denoise the individual's representation, thereby improving the robustness of the regression prediction [36]. Finally, the predicted grade $\hat{y}_i$ is output via an MLP block:

$$\hat{y}_i = g_\phi(h_i^{\text{enr}}) = w_2 \times \text{ReLU}(w_1 h_i^{\text{enr}} + b'_1) + b'_2 \quad (6)$$

where $\phi = \{w_1, w_2, b'_1, b'_2\}$ denotes the proprietary parameters of the regression branch.

3.2. Loss Functions

The training objective of the model comprises three components: regression loss, clustering loss, and prototype regularization, which are combined via a weighted sum to form the total loss. The regression loss employs Mean Squared Error (MSE), defined as follows:

$$\mathcal{L}_{\text{reg}} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (7)$$

This loss function is the standard choice for regression tasks. By calculating the mean of the squared differences between the predicted values $\hat{y}_i$ and the ground-truth values y_i , it provides a clear and differentiable optimization target for the model, effectively penalizing large prediction errors and guiding the model to learn an accurate mapping for grade prediction.

The clustering loss is designed based on contrastive learning principles, aiming to guide the feature representations to form discriminative clustering structures within the latent space. We adopt a variant of the InfoNCE loss:

$$\mathcal{L}_{\text{clu}} = -\frac{1}{n} \sum_{i=1}^n \log \frac{\exp(s_{i,c_i}/\tau)}{\sum_{j=1}^k \exp(s_{i,j}/\tau)} \quad (8)$$

where s_{ij} denotes the cosine similarity between the sample feature and the j -th cluster prototype in the latent space, and $c_i = \arg \max_j s_{ij}$ is the hard-assigned pseudo-label for sample i . This loss function seeks to maximize the mutual information between the sample feature and its assigned prototype (positive pair) while minimizing its association with

other unassigned prototypes (negative pairs) [34,35]. From a geometric perspective of the feature space, this optimization process not only significantly enhances the intra-class compactness of homogeneous sample representations but also effectively widens the inter-class separation margin between heterogeneous groups, thereby driving the network to adaptively excavate and characterize the latent manifold structures inherent in the student data under an unsupervised paradigm.

To alleviate the prototype degradation problem common in unsupervised contrastive learning (i.e., multiple learnable prototypes collapsing into the same local region of the feature space), we introduce a cosine similarity-based prototype dispersion regularization term:

$$\mathcal{L}_{\text{reg-p}} = \frac{1}{k(k-1)} \sum_{i=1}^k \sum_{j \neq i}^k \exp\left(\frac{p_i \times p_j}{\|p_i\| \|p_j\|}\right) \quad (9)$$

This regularization term operates within the same cosine metric space as the contrastive clustering loss. By calculating and penalizing the exponential of the cosine similarity between any two distinct prototypes p_i and p_j , it explicitly encourages all cluster centers to maintain strong repulsion within the latent feature space [37]. This ensures that the learned prototypes evenly cover different regions of the feature space, preventing a large number of student samples from being erroneously collapsed and assigned to a single prototype.

Ultimately, the model's total loss function is the weighted sum of the aforementioned three terms:

$$\mathcal{L} = \mathcal{L}_{\text{reg}} + \alpha \mathcal{L}_{\text{clu}} + \beta \mathcal{L}_{\text{reg-p}} \quad (10)$$

where α and β are two adjustable hyperparameters used to balance the relative importance of clustering learning, prototype dispersion, and the core regression task within the overall optimization objective. By jointly optimizing this unified total loss, the model synchronously improves the accuracy of Composite-Score Estimation and the quality of student profiling (clustering), achieving synergy and mutual enhancement between the two tasks [38].

3.3. Training Optimization

The model is trained using an alternating optimization strategy, and the algorithmic workflow is presented in Algorithm 1. A dynamic weight adjustment strategy is employed during training, enabling the model to focus on different tasks at various training stages. Specifically:

$$\alpha(t) = \max\left(\alpha_{\min}, \alpha_{\text{init}} - (\alpha_{\text{init}} - \alpha_{\min}) \times \frac{t}{T}\right) \quad (11)$$

where α_{init} is the initial weight for the clustering loss, $\alpha_{\min}$ is the lower bound constraint for weight decay, t is the current training epoch, and T is the total number of training epochs. In our experiments, we set $\alpha_{\text{init}} = 0.1$ and $\alpha_{\min} = 0.05$. This strategy allows the clustering loss weight to smoothly and linearly decay from 0.1 to 0.05 over the entire training cycle. Consequently, the model pays more attention to the clustering task in the early stages of training to learn population feature structures and smoothly transitions to focusing primarily on optimizing regression prediction accuracy in the later stages. Inspired by the research on multi-task adaptive weighting by Kendall et al. [39], we designed this linear annealing strategy based on the training progression, which effectively mitigates the task competition and gradient imbalance issues commonly encountered in multi-task learning.

Algorithm 1 Training Algorithm for the Multi-Task Student Profiling and Composite-Score Estimation Model

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, number of clusters k , learning rate η , training epochs T , batch size B

Ensure: Trained model parameters θ, ϕ and cluster prototypes P

- 1: Initialize model parameters θ, ϕ and cluster prototypes P
- 2: **for** $t = 1$ to T **do**
- 3: Randomly sample a batch $\mathcal{B} = \{(x_b, y_b)\}_{b=1}^B$ from the training set
- 4: Forward propagation: compute shared representation $h_b = f_\theta(x_b)$
- 5: Compute clustering similarity s_{bj} and assignment probability γ_{bj}
- 6: Compute cluster context h_b^{ctx} and enriched representation h_b^{enr}
- 7: Compute regression prediction $\hat{y}_b = g_\phi(h_b^{\text{enr}})$
- 8: Compute loss: $\mathcal{L} = \mathcal{L}_{\text{reg}} + \alpha(t)\mathcal{L}_{\text{clu}} + \beta\mathcal{L}_{\text{reg-p}}$
- 9: Backpropagation, compute gradients $\nabla_{\theta, \phi, P}\mathcal{L}$
- 10: Update parameters using Adam optimizer: $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$, $\phi \leftarrow \phi - \eta\nabla_\phi\mathcal{L}$, $P \leftarrow P - \eta\nabla_P\mathcal{L}$
- 11: **if** Validation set performance no longer improves **then**
- 12: Early stopping
- 13: **end if**
- 14: **end for**

4. Experiments

To comprehensively evaluate the effectiveness of the proposed multi-task learning framework, systematic experimental validations were designed in this section. We first detail the experimental dataset and its preprocessing pipeline, followed by elucidating the network implementation details, parameter configurations, and evaluation metric system. Finally, we introduce the baseline models used for comparison to ensure experimental rigor and reproducibility.

4.1. Dataset

This study utilizes an authentic teaching dataset from an English course during the 2024–2025 academic year at a university. The dataset covers multidimensional learning performance records of 429 students over an entire semester. All data used in this study were collected from routine course assessment records and were anonymized before analysis. Personally identifiable information was removed, and the dataset was used only for academic research purposes. The analysis was conducted at the aggregate level and did not affect students' course grades or learning opportunities.

The final overall academic grade for this course is a weighted combination of standard regular grades (formative assessment, 50% weight) and the final exam grade (summative assessment, 50% weight). The regular grades are further subdivided into four evaluation dimensions: classroom performance (accounting for 30% of the regular grade), homework scores (30%), project presentations (20%), and autonomous learning (20%). By integrating these five core assessment dimensions (four formative assessments plus one summative assessment), the dataset provides a rich informational basis for panoramically depicting the dynamic learning process of students. The specific feature dimensions and their descriptions are presented in Table 1.

To ensure the reliability and generalization capability of model evaluation, the dataset was strictly partitioned into training, validation, and testing sets. Specifically, 70% of the data was used for model training, 15% for hyperparameter tuning, and the remaining 15% for final independent performance evaluation. To maintain consistency in data distribution before and after partitioning, a stratified sampling strategy based on the distribution of students' overall grades was adopted, effectively avoiding statistical biases introduced by the splitting process. The specific partitioning details are shown in Table 2. The statistical

characteristics, such as the mean grade and standard deviation, of each data subset remain highly consistent with the overall dataset.

Table 1. Dataset Feature Description.

Feature Name	Description
Classroom Performance	Score based on classroom participation and answering questions
Homework Score	Comprehensive score for the quality of regular homework assignments
Project Presentation	Comprehensive score for the mid-term group project presentation
Autonomous Learning	Autonomous learning score on the online learning platform
Final Exam Score	Final written exam score for summative assessment

Table 2. Dataset Split Statistics.

Data Subset	Number of Samples	Ratio (%)	Mean Grade $\pm$ Std Dev
Training Set	300	70.0	76.3 $\pm$ 12.5
Validation Set	64	15.0	75.8 $\pm$ 11.9
Test Set	65	15.0	77.2 $\pm$ 12.1
Total	429	100.0	76.4 $\pm$ 12.3

4.2. Experimental Settings

4.2.1. Data Preprocessing

To guarantee training stability and convergence efficiency, all numerical features and regression targets were standardized. The Z-score method, as shown in Equation (12), was applied to transform the raw data into a distribution with a mean of 0 and a standard deviation of 1. During the prediction phase, an inverse standardization is applied to the model's outputs to restore the original grading scale.

$$x' = \frac{x - \mu}{\sigma} \quad (12)$$

4.2.2. Network Implementation Details

The model was implemented utilizing the PyTorch 2.6.0 framework. Its core architecture consists of a Shared Feature Extractor, a Clustering Branch, and a Regression Branch. To fully exploit the potential of process-oriented data, when practically constructing the network inputs, we losslessly expanded the 5 macroscopic evaluation dimensions from Table 1 into 10 fine-grained raw observational features to serve as inputs for the shared encoder. Specifically, these 10 independent metrics strictly correspond to the underlying raw records on the teaching platform, including classroom performance; Pigaiwang 1, Pigaiwang 2, and network tests under the homework dimension; mid-term grades under the project presentation dimension; ismart and New Horizon under the autonomous learning dimension; two extracurricular bonus point items (participation bonus and award bonus); and the final exam score as the summative assessment. Directly inputting foundational raw scores effectively avoids information decay and data leakage risks that may arise from manual aggregation.

To prevent overfitting on small-sample educational data, we designed a lightweight network architecture. The Shared Feature Extractor employs a two-layer fully connected network, with each layer followed by Batch Normalization, a ReLU activation function, and a Dropout layer (dropout rate set to 0.1 to preserve weak feature signals). The Clustering Branch achieves contrastive clustering by learning a set of optimizable prototype vectors, while the Regression Branch utilizes cluster context information for feature en-

hancement, ultimately completing precise prediction of overall grades through a regression tower. Detailed parameter configurations are provided in Table 3.

Table 3. Model Architecture Parameter Configuration.

Component	Parameter	Value
Shared Encoder	Input Dimension	10
	Hidden Dimension	64
	Output Dimension	32
	Dropout Rate	0.1
Clustering Branch	Number of Prototypes k	4
	Temperature Parameter τ	0.5
Regression Branch	Input Dimension	32
	Hidden Dimension	32
	Output Dimension	1
	Dropout Rate	0.1
Feature Fusion	Fusion Coefficient λ	0.1
Regularization	Prototype Dispersion Coeff β	0.02

4.2.3. Training Parameters

Model training utilized the Adam optimizer with an initial learning rate set to 5×10^{-3} . Weight Decay (1×10^{-3}) was introduced to provide L2 regularization and enhance model robustness. The training batch size was 32, and the maximum number of epochs was 200. An early stopping mechanism was employed, terminating training prematurely if the validation loss failed to decrease for 25 consecutive epochs. All experiments were run with a fixed random seed (42) for reproducibility. Hyperparameters, including the number of prototypes k , temperature τ , fusion coefficient λ , and loss weights α_{init} and β , were selected via grid search on the validation set. The final values are reported in Table 3. For the baseline models (XGBoost, RF, SVR, LSTM), default hyperparameter settings from their respective libraries were used, and for XGBoost and RF, a brief manual tuning of the number of estimators and maximum depth was performed to ensure fair comparison. All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4070 Ti Super GPU (16 GB VRAM). Training the MTL-Frame model for 200 epochs completes within approximately one minute with early stopping; the slower LSTM baseline requires a few minutes. To effectively balance the synergy and competition between the dual tasks, the clustering loss weight α adopted a dynamic decay strategy: it was initialized to 0.1 and linearly decayed to 0.05 as the epochs increased. This adaptive weighting mechanism successfully prevents excessive interference (negative transfer) from the clustering objective on regression features, ensuring the model balances clustering structure learning early in training, and highly focuses on optimizing regression accuracy in later stages.

4.2.4. Evaluation Metrics

To comprehensively assess the overall performance of the proposed multi-task learning framework, we selected appropriate standardized evaluation metrics for both the composite-score estimation and student profiling tasks.

For the Composite-Score Estimation (regression task), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) were chosen as core evaluation metrics. Specifically, RMSE and MAE quantify the residuals between predicted scores and actual grades from the perspectives of error amplification scale and absolute deviation, respectively. The R^2 metric is utilized to measure the proportion of total

variance in academic performance explained by the model. These three metrics are mutually complementary, comprehensively reflecting the predictive accuracy and goodness-of-fit of the regression model. Given that these three metrics are standard and widely applied in regression tasks, their specific calculation formulas are omitted here.

For the student profiling (clustering task), the Silhouette Score (SC) was adopted as the quantitative evaluation standard for clustering quality. This metric effectively delineates the compactness and boundary clarity of student population segmentations under unsupervised conditions by comprehensively calculating the intra-class cohesion and inter-class separation of samples. For a dataset containing n samples, the overall Silhouette Score is defined as the mean of the Silhouette Scores of all samples, calculated as follows:

$$SC = \frac{1}{n} \sum_{i=1}^n \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (13)$$

where $a(i)$ denotes the average Euclidean distance between sample i and all other samples within its own cluster, quantifying intra-class cohesion; $b(i)$ denotes the average Euclidean distance between sample i and all samples in its nearest neighbor cluster (i.e., the cluster with the minimum average distance excluding its own), representing inter-class separation. The value of the Silhouette Score SC ranges from $[-1, 1]$. Values closer to 1 indicate stronger intra-class compactness and higher inter-class separation, signifying more optimal clustering results.

4.3. Baseline Models

To holistically evaluate the efficacy of the proposed framework, we selected representative baseline models for comparison across the two independent dimensions: composite-score estimation (regression) and student profiling construction (clustering).

For the regression prediction task, typical models representing different paradigms were chosen: Extreme Gradient Boosting (XGBoost) [40] as a representative state-of-the-art non-linear ensemble tree model; Random Forest (RF) [41] representing Bagging ensemble models with strong noise resistance; standard Support Vector Regression (SVR) [42] as a representative kernel method for handling small-sample data; and Long Short-Term Memory networks (LSTM) [43] to assess the performance of complex deep neural networks on the Composite-Score Estimation task. Additionally, to represent modern deep learning approaches explicitly designed for tabular data, we included TabNet [44], which utilizes sequential attention for feature selection, and FT-Transformer [45], a robust self-attention-based architecture adapted from the Transformer model for tabular domains.

For the clustering profiling task, four baseline models encompassing diverse paradigms were selected: the K-Means algorithm [46] representing distance-based traditional clustering; Gaussian Mixture Models (GMMs) [47] representing probability distribution-based soft clustering; Spectral Clustering (SC) [48], which captures manifold structures in data via graph-theoretic methods; and Deep Embedded Clustering (DEC) [49], acting as a representative of deep unsupervised learning by leveraging autoencoders and KL divergence optimization to learn deep clustering representations. By comparing our framework against these single-task models, we aim to validate the advantages of the proposed multi-task framework in achieving collaborative task optimization.

5. Results and Analysis

This section systematically presents and analyzes the experimental results of the proposed multi-task learning framework on the East China Jiaotong University English course dataset. First, we validate the superiority of the proposed framework relative to multiple baseline models through comparative experiments. Second, we deeply dissect the

contributions and operational mechanisms of each core component via ablation studies. Finally, we explore the impact of critical hyperparameters on model performance through sensitivity analysis.

5.1. Comparative Experiments

To comprehensively evaluate the effectiveness of the proposed multi-task learning framework (MTL-Frame), we conducted a comprehensive comparison against the aforementioned eight single-task baseline models across the two categories. Tables 4 and 5 detail the performance of each model on the regression and clustering tasks, respectively.

Table 4. Performance comparison for regression tasks. 5-fold cross-validation results (CV RMSE, CV MAE) are reported.

Model	RMSE	MAE	R ²	CV RMSE	CV MAE
XGBoost	2.8340	1.1213	0.9508	2.91 ± 0.35	1.18 ± 0.15
SVR	8.7188	2.6013	0.5341	8.92 ± 1.14	2.71 ± 0.28
RF	4.5964	2.4038	0.8705	4.81 ± 0.62	2.55 ± 0.31
LSTM	5.1513	1.0161	0.8374	5.34 ± 0.89	1.12 ± 0.19
TabNet	3.1245	1.2560	0.9385	3.25 ± 0.42	1.31 ± 0.18
FT-Transformer	2.9512	1.1845	0.9452	3.08 ± 0.38	1.22 ± 0.16
MTL-Frame	1.7532	0.8168	0.9812	1.79 ± 0.21	0.83 ± 0.09

Table 5. Performance comparison for clustering tasks (Silhouette Score).

Model	Silhouette Score
K-Means	0.2555
GMM	0.2511
SC	0.4912
DEC	0.6165
MTL-Frame	0.3562

The experimental results demonstrate that the proposed MTL-Frame achieved state-of-the-art performance in the regression prediction task. Specifically, MTL-Frame attained an RMSE of 1.7532, an MAE of 0.8168, and an R² of 0.9812, comprehensively outperforming advanced ensemble tree models like XGBoost and Random Forest (RF), as well as the deep learning baseline LSTM. Furthermore, compared with modern deep learning architectures designed for tabular data, TabNet and FT-Transformer show competitive performance but fail to surpass XGBoost and significantly underperform compared to MTL-Frame (e.g., FT-Transformer achieved an RMSE of 2.9512). This aligns with the recent literature indicating that modern deep tabular models are often highly data-hungry and prone to overfitting on small-sample datasets ($n = 429$). This profoundly indicates that by incorporating an unsupervised population clustering branch, MTL-Frame successfully captures the latent distributional structures of student populations. This population prior knowledge acts as powerful contextual regularization, dramatically suppressing the overfitting tendencies of neural networks on small-sample educational tabular data and achieving significant positive task transfer.

In the student population profiling (clustering) task, the Silhouette Score of MTL-Frame reached 0.3562, significantly outperforming traditional K-Means (0.2555) and GMM (0.2511) algorithms, which proves the dimensionality reduction advantage of deep feature extraction when processing fine-grained, multi-source data. Although this metric is lower than that of SC (0.4912) and DEC (0.6165)—models designed exclusively for pure unsupervised clustering—this phenomenon perfectly aligns with the trade-off mechanism inherent in multi-task learning. Single-task models like DEC aim solely to maximize geometric separation in the feature space as their single optimization objective. Conversely, MTL-Frame

mandates that the clustering space must serve downstream precise composite-score estimation. Consequently, MTL-Frame sacrifices localized geometric separation in exchange for peak performance in global regression accuracy. This demonstrates that the profiles generated by MTL-Frame represent semantic clustering driven by the strong supervision of authentic academic outcomes, endowing them with stronger pedagogical interpretability and practical decision-support value compared to purely geometric clustering.

To assess the statistical significance of the improvements, we conducted 5-fold cross-validation. The proposed MTL-Frame achieved an average RMSE of 1.79 ± 0.21 and MAE of 0.83 ± 0.09 , while the best baseline (XGBoost) obtained 2.91 ± 0.35 and 1.18 ± 0.15 , respectively. A paired t-test confirms that MTL-Frame significantly outperforms XGBoost ($p < 0.01$). The cross-validation results for all regression models are reported directly in Table 4.

5.2. UMAP Visualization of Learned Representations

To qualitatively evaluate the clustering structure, we projected the shared representations h of the test samples into two dimensions using UMAP 0.5.3. Figure 2 displays the resulting embedding, where points are colored by their assigned cluster. The visualization reveals four distinct, albeit partially overlapping, groups, confirming that the prototype-based branch captures meaningful population segmentation. The moderate overlap explains the moderate Silhouette Score and is consistent with the multi-task trade-off: the regression task pulls together samples with similar academic outcomes, even if they belong to different behavioral clusters. Brief characterization of the four clusters based on centroid features suggests: Cluster 1 (high overall achievers) exhibits strong performance across all dimensions, representing well-rounded students who excel in both formative and summative assessments; Cluster 2 (strong autonomous learners) scores highest on iSmart V2.3.0 and New Horizon 0.5.2 platforms, indicating self-directed learning habits but with moderate exam scores; Cluster 3 (homework-dependent students) relies heavily on Pigaiwang assignments and network tests for their overall grade, showing diligence in regular tasks but weaker autonomous learning; and Cluster 4 (classroom participators with lower exam scores) demonstrates active in-class engagement but struggles in final examinations, suggesting a need for exam preparation support. These profiles provide actionable guidance for differentiated instructional interventions.

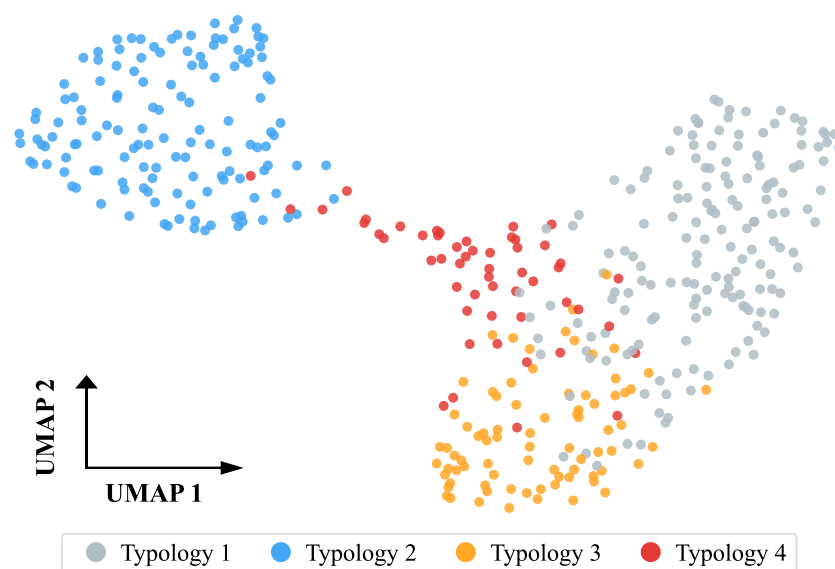


Figure 2. UMAP visualization of the learned student representations, colored by assigned cluster. Four groups with clear trends are visible, demonstrating the profiling capability of MTL-Frame.

5.3. Model Interpretability via SHAP Analysis

To address the interpretability requirement, we applied SHapley Additive exPlanations (SHAP) to the regression branch. Figure 3 shows the mean absolute SHAP values for the 10 input features. As shown, while the Final Exam Score understandably holds the highest feature importance, formative assessment indicators—particularly Autonomous Learning and Project Presentation—exhibit highly competitive SHAP values. This robust feature distribution unequivocally demonstrates that MTL-Frame does not merely reconstruct the final grade from the summative exam, but effectively mines deep behavioral patterns from the students' learning process.

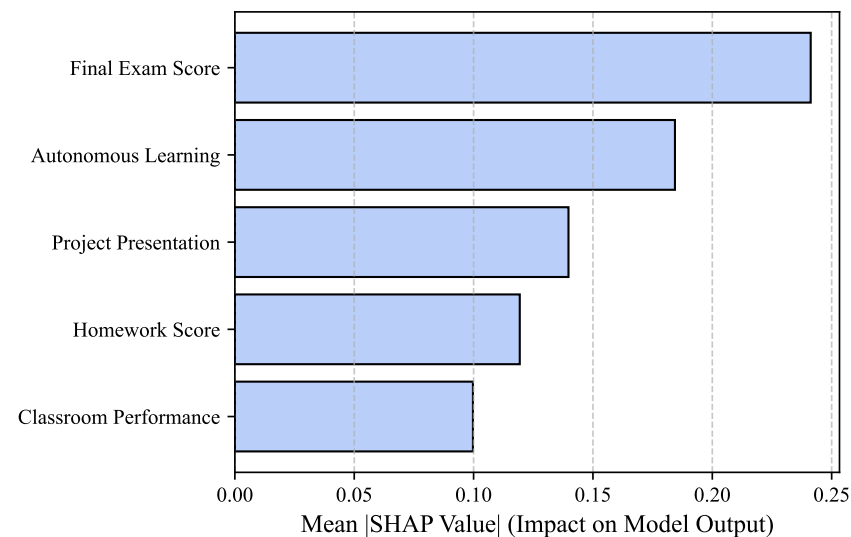


Figure 3. SHAP feature importance plot. Features are ordered by mean $|\text{SHAP}|$.

5.4. Ablation Study

To deeply understand the contributions of each core component within the proposed framework, this section designs a systematic ablation study. We constructed four variant models by independently removing the clustering branch, the regression branch, the prototype regularization, and the dynamic weight adjustment mechanism for comparative analysis.

5.4.1. Effectiveness of Core Components

Table 6 illustrates the performance evolution of the models after removing respective core components. Here, “w/o Clustering” denotes the removal of the clustering branch; “w/o Regression” signifies the removal of the regression branch; “w/o Prototype Reg” indicates the omission of the prototype dispersion regularization term; “w/o Dynamic Weight” denotes the use of fixed loss weights; and “w/o Exam Score” refers to the exclusion of the final examination score from the input features.

Table 6. Ablation Study Results of the Core Components in the Multi-Task Learning Framework.

Model Variant	RMSE	MAE	R ²	Silhouette Score
w/o Clustering	4.0904	1.3360	0.8975	—
w/o Regression	—	—	—	0.6218
w/o Prototype Reg	1.6196	0.9584	0.9839	0.2301
w/o Dynamic Weight	2.7905	0.9360	0.9523	0.5319
w/o Exam Score	2.3167	1.1283	0.9681	0.3423
MTL-Frame	1.7532	0.8168	0.9812	0.3562

The experimental outcomes confirm that all core components contribute significantly to final performance. When the clustering branch is removed (w/o Clustering), the model exhibits a conspicuous overfitting tendency on the small-sample dataset, with RMSE surging significantly from 1.7532 to 4.0904. This validates that the population prior knowledge provided by unsupervised clustering serves a critical contextual regularization function, assisting the regression task in escaping local optima.

Upon removing the regression branch (w/o Regression), the Silhouette Score rises to 0.6218. This illustrates the performance trade-off in multi-task learning: the strong supervisory signals of the regression task constrain the geometric distribution of the feature space. While this reduces local inter-class separation, it endows the profile with enhanced pedagogical interpretability.

The absence of the prototype regularization term (w/o Prototype Reg) causes the Silhouette Score to plummet to 0.2301. This phenomenon indicates that explicit prototype repulsion constraints are vital for preventing representation collapse. Furthermore, the removal of the dynamic weight adjustment mechanism (w/o Dynamic Weight) results in substantial RMSE fluctuations, demonstrating the necessity of adaptive loss balancing strategies in mitigating task competition and gradient dominance issues.

A particularly instructive variant is the removal of the final examination score (w/o Exam Score). Even without this single most predictive feature, MTL-Frame still achieves an RMSE of 2.3167 and an R^2 of 0.9681, which remains markedly superior to XGBoost (RMSE 2.8340, R^2 0.9508) trained on the full 10-feature set, despite MTL-Frame having access to one fewer feature. This finding directly confirms that the model extracts meaningful predictive signals from the formative assessment indicators and the learned student-group structure, rather than relying on a trivial reconstruction of the scoring formula. Notably, the Silhouette Score remains virtually unchanged (0.3423 vs. 0.3562), indicating that the clustering structure is largely preserved even without the exam score, which further underscores that the multi-task profiling is driven by holistic behavioral patterns rather than by a single dominant feature. The moderate increase in error relative to the full model is expected, yet the maintained high R^2 and stable clustering quality jointly demonstrate the robustness of the joint learning paradigm.

5.4.2. Sensitivity Analysis on the Number of Clusters

The number of clusters k is a crucial hyperparameter affecting the quality of model feature enhancement. To investigate its impact, we conducted experiments with k ranging from 2 to 7. The results are presented in Table 7. At $k = 4$, the model achieves an optimal balance between the regression and clustering tasks, with all evaluation metrics in their most favorable state. A k that is too small leads to overly coarse population segmentation, failing to provide high-resolution auxiliary features; conversely, a k that is too large results in fragmented segmentation, introducing redundant noise and degrading prediction accuracy.

Table 7. Model Performance under Varying Number of Clusters k (Bolded data represent the optimal parameter values).

k	RMSE	MAE	R^2	Silhouette Score
2	2.2695	0.8717	0.9684	0.2153
3	2.7230	1.0580	0.9546	0.2327
4	1.7532	0.8168	0.9812	0.3562
5	2.9725	1.0890	0.9458	0.2690
6	2.9933	0.9347	0.9451	0.3293
7	2.1421	0.9300	0.9719	0.3103

5.4.3. Sensitivity Analysis on Loss Weights

The initial value of the loss weight α_{init} and the prototype penalty coefficient β respectively govern the relative importance of the clustering task in the overall optimization objective and the dispersion degree of clustering prototypes in the feature space. Table 8 reports the dual-task evaluation results based on various parameter combinations. It is evident that selecting an appropriate parameter combination is necessary to obtain high-quality multi-task representations, and $\alpha_{init} = 0.10$ with $\beta = 0.02$ likely represents the optimal configuration balancing Composite-Score Estimation accuracy and population profiling quality.

Table 8. Impact of Different Loss Weight Combinations on Model Performance (Bolded data represent the optimal parameter values).

Initial α	β	RMSE	MAE	Silhouette Score
0.10	0.01	1.9534	0.8879	0.3721
0.10	0.02	1.7532	0.8168	0.3562
0.10	0.03	1.4254	0.8683	0.3423
0.07	0.01	2.0557	0.8322	0.3322
0.07	0.02	2.0942	0.8886	0.3255
0.07	0.03	2.1839	0.8772	0.3412
0.05	0.01	1.6708	0.7353	0.3011
0.05	0.02	2.3281	0.9019	0.2880
0.05	0.03	2.5068	0.8973	0.2579

Although the model with $\alpha_{init} = 0.05$ and $\beta = 0.01$ achieved a very low Mean Absolute Error (MAE of 0.7353), its Silhouette Score noticeably declined (0.3011), suggesting degradation in the unsupervised clustering structure. Similarly, over-weakening the prototype repulsion (e.g., $\alpha_{init} = 0.10, \beta = 0.01$) achieves the highest Silhouette Score (0.3721) but leads to a clear increase in Root Mean Square Error (RMSE). This indicates that the model exhibits a certain sensitivity to task competition, and rational loss weight settings are conducive to mitigating this issue.

5.4.4. Analysis of Fusion Coefficient λ

The fusion coefficient λ controls the intervention intensity and relative importance of clustering context information on the primary regression task. Table 9 reports the multi-task joint evaluation results based on varying λ values. It is clear that selecting a suitable λ is crucial for the model to acquire high-quality, dual-task synergistic representations. A value of $\lambda = 0.10$ is likely the optimal parameter configuration for balancing composite-score estimation accuracy with population profiling quality.

Table 9. Impact of Different Feature Fusion Coefficients λ on Model Performance (Bolded data represent the optimal parameter values).

λ	RMSE	MAE	R ²	Silhouette Score
0.01	1.8810	0.8582	0.9783	0.3590
0.05	1.9588	0.9127	0.9765	0.3108
0.10	1.7532	0.8168	0.9812	0.3562
0.15	1.1854	0.7403	0.9914	0.2060
0.30	2.4612	0.9689	0.9629	0.2877

Although the model with $\lambda = 0.15$ attained extreme performance in regression prediction (RMSE plunging to 1.1854), its clustering Silhouette Score plummeted significantly to 0.2060. This indicates that excessive feature information injection triggers severe negative transfer and representation collapse, disrupting the unsupervised manifold structure. Con-

versely, an excessively small λ (such as 0.01) renders it difficult for the model to effectively leverage the synergistic enhancement of multiple tasks.

6. Discussion

This study set out to investigate whether the joint optimization of unsupervised student profiling and supervised grade estimation can yield synergistic benefits over traditional single-task pipelines. The results clearly demonstrate the effectiveness of MTL-Frame, which achieved substantial improvements in regression accuracy while maintaining interpretable student groupings. Below, we discuss the implications of these findings, the trade-offs inherent in multi-task learning, the limitations of the current work, and the broader practical significance.

Interpretation of Performance Gains and Clustering Trade-off. The marked reduction in RMSE (from 2.8340 for XGBoost to 1.7532) validates the core hypothesis: the prototype-based clustering branch acts as a contextual regularizer, injecting population-level priors into the individual prediction and effectively combating overfitting in the small-sample regime. The SHAP analysis further corroborates that the model attends to educationally relevant features, with homework and autonomous learning indicators playing dominant roles, consistent with prior research emphasizing process-oriented assessment [14]. Conversely, MTL-Frame's Silhouette Score (0.356) is lower than that of pure clustering methods like DEC (0.616). This outcome is an intrinsic consequence of the multi-task objective: while pure clustering maximizes geometric separation, MTL-Frame must simultaneously accommodate the regression loss, which groups students with similar outcome scores even if their behavioral patterns differ. Consequently, the resulting clusters are geometrically less separated and their boundaries are blurred compared to those produced by dedicated clustering models. However, this trade-off is pedagogically justified: the profiles are shaped not only by behavioral similarity but also by academic outcome proximity, making them more actionable for identifying at-risk students and informing interventions. The UMAP visualization confirms that the obtained clusters remain semantically meaningful, and the four identified profiles broadly correspond to high achievers, strong autonomous learners, homework-dependent students, and classroom participants with lower exams.

On the Nature of the Regression Task. A clarifying remark is warranted regarding the use of the final examination score as both an input feature and a component of the target variable. This design does not constitute data leakage, for the following reasons. First, the model receives 10 fine-grained input indicators but is not provided with the weighting formula that combines them into the final grade; it must learn this complex, multi-level aggregation function from the data. Second, the final examination score is only one of 10 input dimensions, and the model's architecture inherently prevents over-reliance on any individual feature through shared representation learning and cluster-based contextual regularization. As explicitly demonstrated by the SHAP analysis, process-oriented indicators such as autonomous learning and project presentations carry substantial weight, preventing the model from degenerating into a trivial mapping of the exam score. Third, using formative assessment components to model summative outcomes is an established practice in educational data mining, where the goal is to understand and replicate the assessment logic itself. The reported extremely high accuracy ($R^2 = 0.9812$) thus reflects the model's ability to faithfully learn the course's scoring mechanism, a non-trivial estimation task with practical value for assessment auditing, grade verification, and personalized feedback.

Sensitivity and Stability. The sensitivity analysis revealed that the fusion coefficient λ strongly influences the balance between tasks: at $\lambda = 0.15$, regression accuracy peaks (RMSE = 1.1854) but the Silhouette Score drops sharply to 0.206, indicating a near-complete collapse of clustering structure. This pronounced sensitivity indicates a degree of inherent

instability in the task balance, a recognized challenge in multi-task learning when competing objectives have different convergence dynamics. Our linear annealing strategy for $\alpha(t)$ partially mitigates this issue by gradually reducing the clustering objective's influence over training. More advanced methods such as uncertainty weighting [39] or gradient-based task balancing could further stabilize training and should be explored in future work.

Generalizability, Limitations, and Future Directions. Currently, the framework has been validated only on a single English course at one university, limiting the generalizability of the findings. As pedagogical strategies, grading criteria, and student demographics vary significantly across different institutions and disciplines, the extent to which these findings transfer to other higher-education settings remains an open question. Despite the small sample size (429 students), the lightweight architecture combined with aggressive dropout, batch normalization, and L2 regularization proved sufficient to avoid overfitting, as evidenced by the strong cross-validation results and the substantial performance gap between MTL-Frame and the single-task baseline without clustering (w/o Clustering, RMSE = 4.0904). Nevertheless, extending the evaluation to multiple disciplines, institutions, and cultural contexts is necessary before practical deployment. Beyond the single-course limitation, all features were structured numerical scores; incorporating unstructured data such as forum text or clickstream logs could enrich the profiles. The baseline comparison focuses on widely adopted methods in educational data mining; a systematic evaluation against modern deep learning architectures for tabular data on larger educational datasets is a necessary next step. Additional directions include adaptive cluster determination, instance-level explanations for individual student support, and longitudinal studies on the impact of such models on teaching practice. The demonstrated success of joint learning offers a promising blueprint for building integrated educational AI tools that simultaneously provide accurate performance estimates and rich learner profiles.

7. Conclusions

This paper presented an end-to-end multi-task learning collaborative framework designed to overcome the fragmentation between student profiling construction and composite-score estimation (or assessment auditing) inherent in traditional educational data mining. Experimental results on a real-world, fine-grained educational dataset encompassing 429 students demonstrate that MTL-Frame achieves excellent performance in recovering the overall grade formula (RMSE = 1.7532, MAE = 0.8168, $R^2 = 0.9812$). While this high R^2 partly reflects the model's ability to learn the deterministic scoring aggregation, the performance remains strong even when the final exam score is excluded from the inputs ($R^2 = 0.9681$, RMSE = 2.3167), confirming genuine estimation capability beyond trivial memorization. Simultaneously, the model extracts four semantically meaningful groups with strong pedagogical guidance in the unsupervised clustering task (Silhouette Score of 0.3562). Although this score is lower than those of dedicated clustering models, the resulting profiles are shaped by both behavioral patterns and academic outcomes, yielding educationally actionable segmentations despite somewhat blurred boundaries. Compared to state-of-the-art single-task ensemble models like XGBoost and traditional deep neural networks, MTL-Frame proves that unsupervised population profiling priors can serve as an effective contextual regularization mechanism, significantly mitigating the overfitting phenomena common in deep models applied to small-sample educational tabular data. Exhaustive ablation studies further validate the efficacy and necessity of core modules such as the shared representation mechanism, cluster-aware feature enhancement, and prototype dispersion regularization.

Despite the considerable application potential demonstrated by the proposed framework, the current validation is limited to a single English course at one university, and the

model's effectiveness on courses in technical, medical, or other humanities disciplines remains unverified. In addition, the single-institution setting cannot account for variations in grading policies, student demographics, and instructional designs that exist across different higher-education contexts. Extending the evaluation to multiple disciplines, institutions, and cultural contexts is therefore a critical next step before any practical deployment can be considered. The lightweight and modular nature of MTL-Frame should facilitate such cross-contextual validation, as the architecture can be readily retrained on new datasets with minimal adaptation.

Furthermore, the model inputs lean heavily toward structured, process-oriented evaluation data. Future work will concentrate on three key dimensions: First, constructing cross-disciplinary and cross-institutional benchmark datasets to fully validate the framework's robustness and generalization capabilities across different educational stages and cultural contexts. Second, introducing multimodal fusion mechanisms to integrate unstructured behavioral trajectories—such as online platform clickstreams and classroom interaction text—thereby constructing more panoramic learner profiles. Finally, incorporating eXplainable Artificial Intelligence (XAI) techniques to develop post hoc attribution analysis modules, further unboxing the black box of multi-task collaborative decision-making, and providing more transparent technical support for large-scale, trustworthy personalized interventions in real-world teaching scenarios.

Author Contributions: Conceptualization, G.J.; methodology, G.J.; software, M.T.; validation, M.T.; investigation, G.J.; resources, G.J.; writing—original draft preparation, M.T. and Z.Y.; writing—review and editing, Z.Y.; visualization, M.T.; project administration, Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly funded by the Project of Jiangxi Provincial Education Science No. 1103522018, and the National Foundation of Science of China No. 62463009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EDM	Educational Data Mining
MTL	Multi-Task Learning
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
SC	Silhouette Score
XGBoost	Extreme Gradient Boosting
RF	Random Forest
SVR	Support Vector Regression
LSTM	Long Short-Term Memory
GMM	Gaussian Mixture Model
DEC	Deep Embedded Clustering

References

1. Sosa-Díaz, M.J.; Garrido-Arroyo, M.d.C.; González Delgado, M.Y. Transformation of Educational Models in Higher Education During and After “Emergency Remote Teaching”. *Educ. Sci.* **2025**, *15*, 1249.

2. Romero, C.; Ventura, S. Data mining in education. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 12–27.
3. Siemens, G. Learning analytics: The emergence of a discipline. *Am. Behav. Sci.* **2013**, *57*, 1380–1400.
4. Xie, H.; Chu, H.C.; Hwang, G.J.; Wang, C.C. Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Comput. Educ.* **2019**, *140*, 103599.
5. Desmarais, M.C.; Baker, R.S.d. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Model. User-Adapt. Interact.* **2012**, *22*, 9–38.
6. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355.
7. Ren, Y.; Pu, J.; Yang, Z.; Xu, J.; Li, G.; Pu, X.; Yu, P.S.; He, L. Deep clustering: A comprehensive survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 5858–5878.
8. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609.
9. Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. A systematic review of deep learning approaches to educational data mining. *Complexity* **2019**, *2019*, 1306039.
10. Molnar, C. *Interpretable Machine Learning*; Lulu Press: Morrisville, NC, USA, 2020.
11. Zhao, S.; Sahebi, S. Multi-task modeling of student knowledge and behavior. In *33rd ACM International Conference on Information and Knowledge Management*; ACM Digital Library: New York, NY, USA, 2024; pp. 3363–3373.
12. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
13. Di Mitri, D.; Schneider, J.; Specht, M.; Drachsler, H. From signals to knowledge: A conceptual model for multimodal learning analytics. *J. Comput. Assist. Learn.* **2018**, *34*, 338–349.
14. Gašević, D.; Dawson, S.; Siemens, G. Let's not forget: Learning analytics are about learning. *TechTrends* **2015**, *59*, 64–71.
15. Khosravi, H.; Shum, S.B.; Chen, G.; Conati, C.; Tsai, Y.S.; Kay, J.; Knight, S.; Martinez-Maldonado, R.; Sadiq, S.; Gašević, D. Explainable artificial intelligence in education. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100074.
16. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.
17. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215.
18. Alamgir, Z.; Akram, H.; Karim, S.; Wali, A. Enhancing student performance prediction via educational data mining on academic data. *Inform. Educ.* **2024**, *23*, 1–24.
19. Shawareb, N.; Ewais, A.; Dalipi, F. *Utilizing Data Mining Techniques to Predict Students Performance Using Data Log from MOODLE*; Arab American University: Jenin, Palestine, 2024.
20. Lai, X.; Hao, W.; Zhang, Z.; Bai, X. Predicting student performance with graph structure density-based graph neural networks. In *Machine Learning and Intelligent Computing*; PMLR: New York, NY, USA, 2024; pp. 25–32.
21. Alyahyan, E.; Düşteğör, D. Predicting academic success in higher education: Literature review and best practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3.
22. Al-Tameemi, G.; Xue, J.; Ali, I.H.; Ajit, S. A hybrid machine learning approach for predicting student performance using multi-class educational datasets. *Procedia Comput. Sci.* **2024**, *238*, 888–895.
23. Bharara, S.; Sabitha, S.; Bansal, A. Application of learning analytics using clustering data Mining for Students' disposition analysis. *Educ. Inf. Technol.* **2018**, *23*, 957–984.
24. Song, R.; Pang, F.; Jiang, H.; Zhu, H. A machine learning based method for constructing group profiles of university students. *Heliyon* **2024**, *10*, e29181.
25. Salles, I.; Mejia-Domenzain, P.; Swamy, V.; Blackwell, J.; Käser, T. Interpret3C: Interpretable student clustering through individualized feature selection. In *International Conference on Artificial Intelligence in Education*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 382–390.
26. Kizilcec, R.F.; Piech, C.; Schneider, E. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Third International Conference on Learning Analytics and Knowledge*; ACM Digital Library: New York, NY, USA, 2013; pp. 170–179.
27. Tiukhova, E.; Vemuri, P.; Óskarsdóttir, M.; Poelmans, S.; Baesens, B.; Snoeck, M. *Discovering Unusual Study Patterns Using Anomaly Detection and XAI*; University of Hawaii System: Honolulu, HI, USA, 2024.
28. Huang, T.; Hu, S.; Yang, H.; Geng, J.; Li, Z.; Xu, Z.; Ou, X. Response speed enhanced fine-grained knowledge tracing: A multi-task learning perspective. *Expert Syst. Appl.* **2024**, *238*, 122107.
29. Geden, M.; Emerson, A.; Rowe, J.; Azevedo, R.; Lester, J. Predictive student modeling in educational games with multi-task learning. In *AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2020; Volume 34, pp. 654–661.
30. Conijn, R.; Snijders, C.; Kleingeld, A.; Matzat, U. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Trans. Learn. Technol.* **2016**, *10*, 17–29.

31. Chrysafiadi, K.; Virvou, M. Student modeling approaches: A literature review for the last decade. *Expert Syst. Appl.* **2013**, *40*, 4715–4729.
32. Baashar, Y.; Alkawsi, G.; Mustafa, A.; Alkahtani, A.A.; Alsariera, Y.A.; Ali, A.Q.; Hashim, W.; Tiong, S.K. Toward predicting student's academic performance using artificial neural networks (ANNs). *Appl. Sci.* **2022**, *12*, 1289.
33. Vandenhende, S.; Georgoulis, S.; Van Gansbeke, W.; Proesmans, M.; Dai, D.; Van Gool, L. Multi-task learning for dense prediction tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3614–3633.
34. Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J.T.; Peng, X. Contrastive clustering. In *AAAI conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2021; Volume 35, pp. 8547–8555.
35. Li, J.; Zhou, P.; Xiong, C.; Hoi, S.C. Prototypical contrastive learning of unsupervised representations. *arXiv* **2020**, arXiv:2005.04966.
36. Blikstein, P.; Worsley, M. Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *J. Learn. Anal.* **2016**, *3*, 220–238.
37. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *arXiv* **2017**, arXiv:1703.05175.
38. Shen, J.; Wang, Q.; Xiao, Z.; Van Noord, N.; Worring, M. Go4align: Group optimization for multi-task alignment. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 111382–111405.
39. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 7482–7491.
40. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*; ACM Digital Library: New York, NY, USA, 2016; pp. 785–794.
41. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
42. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.
43. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
44. Arik, S.O.; Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *arXiv* **2020**, arXiv:1908.07442.
45. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*; ACM Digital Library: New York, NY, USA, 2021; Volume 34, pp. 18932–18943.
46. McQueen, J.B. Some methods of classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*; University of California, Berkeley: Berkeley, CA, USA, 1967; pp. 281–297.
47. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; John Wiley & Sons: Hoboken, NJ, USA, 2000.
48. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395–416.
49. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2016; pp. 478–487.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.