Toward Adaptive Educational Intervention: Meta-Adaptive Cross-Modal Gating for Few-Shot Personalized Intervention

Houda Kaa, Hanane Allioui, Ilham Oumaira
Laboratory: Engineering Sciences, Ibn Tofail University, Morocco

***Abstract*—The rapid evolution of AI-enhanced learning environments has created an urgent need for intelligent educational systems capable of delivering early and personalized interventions under severe data sparsity conditions. This study proposes a Meta-Adaptive Cross-Modal Gating (MACMG) mechanism integrated within a Multimodal Transformer-based Educational Digital Twin framework for early detection of at-risk students. The proposed approach addresses the cold-start problem by introducing a student-specific, meta-learned gating policy that dynamically fuses textual, behavioral, and physiological educational signals. At the core of MACMG lies a bilevel optimization strategy inspired by Model-Agnostic Meta-Learning (MAML), enabling rapid adaptation to unseen learners using only a few initial interaction episodes. The framework generates time-varying modality weights that emphasize informative signals while suppressing noisy channels. To preserve representational capacity and computational efficiency, the adaptive gating mechanism is integrated only into the top two layers of a six-layer multimodal Transformer using a first-order MAML approximation. The personalized representations are propagated toward a risk prediction module and a digital twin state manager responsible for continuously updating learner knowledge states. Experimental results demonstrate that MACMG achieves AUROC scores of 0.821 on StudentLife and 0.834 on DAiSEE, outperforming static multimodal Transformers, conventional fine-tuning approaches, and full-model meta-learning baselines. Furthermore, the proposed framework reduces learner-specific adaptation time to only 0.47 seconds while maintaining robust performance under sparse-data conditions, highlighting its suitability for real-time personalized educational intervention.**

Keywords*—Multimodal transformer; meta-learning; at-risk student detection; personalized learning; educational data mining; adaptive intervention

I. Introduction

The rapid integration of artificial intelligence into educational technology has significantly transformed the landscape of learning analytics and adaptive educational systems [1, 2]. Modern Learning Management Systems (LMSs) continuously generate large volumes of heterogeneous learner interaction data, including clickstream behavior, textual discussions, temporal activity patterns, emotional engagement indicators, and multimodal interaction traces [3, 4]. Among the emerging AI paradigms applied to these environments, Multimodal Transformer architectures have demonstrated a remarkable ability to fuse heterogeneous educational data streams into unified representations of learner states through cross-modal attention mechanisms [5-7]. These architectures enable the modeling of complex temporal dependencies and cross-modal relationships that traditional machine learning approaches fail to capture effectively.

Simultaneously, the concept of Digital Twins has recently emerged as a promising paradigm for adaptive educational intelligence systems [8, 9]. Digital twins, dynamic virtual replicas of learners continuously synchronized with educational interaction streams, aim to model evolving learner states, cognitive progression, engagement dynamics, and academic risk trajectories in real time [8, 10]. By continuously updating learner representations from multimodal educational traces, educational digital twins offer promising opportunities for personalized intervention, adaptive tutoring, and intelligent pedagogical orchestration. However, despite their potential, existing educational digital twin frameworks remain highly dependent on substantial historical interaction data to generate reliable predictions and adaptive recommendations.

This limitation becomes particularly problematic in cold-start educational scenarios involving newly enrolled students for whom only a few initial interactions are available. This challenge corresponds to the well-known cold-start problem frequently encountered in adaptive educational intelligence systems [11]. Most existing educational AI systems require extensive historical data before producing stable predictions, thereby limiting their effectiveness during the earliest and most critical phases of learner disengagement. Yet, timely intervention during these initial stages is often essential for preventing academic failure, emotional disengagement, and dropout propagation.

Several studies have attempted to address sparse-data educational settings through transfer learning, domain adaptation, and meta-learning approaches [12-15]. In these frameworks, a model is typically pre-trained on historical learner populations and subsequently adapted to unseen students using limited support data. Among these approaches, Model-Agnostic Meta-Learning (MAML) has emerged as one of the most influential paradigms for rapid adaptation under few-shot conditions [12].

MAML optimizes model parameters such that only a few gradient updates are required to specialize the model to new tasks or users. Recent advances further demonstrate the growing importance of meta-learning within domain generalization, multimodal adaptation, and personalized intelligent systems [13-15]. However, existing educational meta-learning

approaches adapt either the entire architecture or large subsets of parameters, increasing computational overhead and susceptibility to overfitting under sparse-data conditions. Furthermore, multimodal educational architectures typically rely on static cross-modal fusion policies learned globally from training populations [5].

Recent educational ecosystems are additionally being transformed by generative artificial intelligence technologies, including large language model (LLM)-based tutoring systems, conversational educational agents, and AI-enhanced LMS environments [16, 17]. Consequently, student interaction patterns increasingly extend beyond conventional LMS traces toward AI-mediated cognitive interaction behaviors such as prompt formulation, semantic dependency, iterative reasoning assistance, and AI-assisted problem solving [18-20]. This evolution challenges conventional predictive learning analytics systems that rely solely on static behavioral representations and motivates the need for adaptive multimodal educational intelligence architectures capable of dynamically modeling evolving AI-augmented learning trajectories.

Adaptive gating mechanisms have previously demonstrated strong effectiveness within Mixture-of-Experts (MoE) architectures by dynamically routing information toward specialized computational pathways [21, 22, 23]. Similarly, multimodal gating approaches enable context-dependent weighting of heterogeneous modalities according to their predictive relevance [7, 24, 25]. Nevertheless, these gating policies are optimized globally during training and remain fixed during inference, preventing rapid learner-specific personalization under cold-start educational conditions.

Despite these advances, several important limitations remain. First, most multimodal educational systems rely on static fusion policies that assume identical modality importance across all learners, despite substantial heterogeneity in learning behaviors and risk manifestations. Second, existing meta-learning approaches generally adapt large portions of the underlying architecture, resulting in increased computational complexity and a higher risk of overfitting under extreme few-shot conditions. Third, current educational digital twin frameworks primarily focus on learner-state prediction and monitoring, while providing limited support for adaptive intervention selection. Consequently, there remains a lack of lightweight educational intelligence frameworks capable of simultaneously achieving rapid personalization, adaptive multimodal fusion, and intervention-aware decision support under cold-start conditions. Addressing this research gap constitutes the primary motivation for the proposed Meta-Adaptive Cross-Modal Gating framework.

Based on the aforementioned limitations, the objective of this study is to develop a lightweight and adaptive multimodal educational intelligence framework capable of rapidly personalizing learner representations under cold-start conditions. More specifically, the proposed research seeks to 1) improve early at-risk student detection from limited interaction data, 2) dynamically adapt modality importance according to individual learner characteristics, 3) reduce computational overhead compared with full-model meta-learning approaches, and 4) support personalized pedagogical intervention within educational digital twin environments.

A meta-learning strategy was selected because the targeted educational scenario is characterized by severe data sparsity during the earliest stages of learner interaction. Traditional supervised learning and fine-tuning approaches generally require substantial historical data before producing reliable predictions. In contrast, meta-learning enables the acquisition of transferable adaptation knowledge from previously observed learners and supports rapid personalization from only a few support samples. Furthermore, the proposed approach adapts only the cross-modal gating mechanism rather than the entire Transformer architecture, thereby reducing computational complexity while limiting overfitting risks under few-shot conditions.

To address these limitations, this study introduces the Meta-Adaptive Cross-Modal Gating (MACMG) framework, a lightweight yet highly adaptive personalization mechanism for autonomous educational digital twins.

The proposed framework dynamically personalizes multimodal fusion policies through a meta-learned gating network capable of rapidly adapting to unseen students using only a few initial interaction episodes. Instead of adapting the entire Transformer architecture, the proposed approach treats the gating mechanism itself as a meta-parameter optimized through a bilevel meta-learning objective [12, 23, 26]. The proposed MACMG framework offers several important advantages. First, by adapting only the lightweight gating mechanism rather than the full multimodal architecture, the framework remains computationally efficient and less vulnerable to overfitting under sparse-data conditions. Second, dynamic gating enables individualized modality weighting, allowing the system to emphasize highly informative signals while suppressing noisy or weakly predictive modalities. Third, the adaptive fusion policy is directly coupled with personalized intervention selection, enabling the framework to operate not merely as a predictive analytics system but as an adaptive educational orchestration mechanism within AI-augmented learning ecosystems.

This study makes several important contributions to the field of AI-enhanced educational intelligence. First, it introduces a Meta-Adaptive Cross-Modal Gating (MACMG) mechanism capable of dynamically personalizing multimodal fusion policies under few-shot conditions. Second, the proposed framework employs a bilevel meta-learning optimization strategy that adapts only a lightweight gating network, thereby improving computational efficiency while maintaining strong adaptation capabilities. Third, extensive experiments on two multimodal educational datasets demonstrate superior early risk detection performance and robustness under severe data sparsity when compared with conventional multimodal Transformers, fine-tuning approaches, and full-model meta-learning baselines. Finally, the framework establishes a direct connection between adaptive multimodal fusion and personalized pedagogical intervention selection within educational digital twin environments.

This study represents one of the earliest attempts to integrate meta-learned cross-modal gating, educational digital twins, and few-shot multimodal personalization within a unified educational intelligence framework. More broadly, the proposed framework reflects the ongoing transition from static predictive learning analytics toward adaptive and autonomous educational intelligence systems capable of continuously modeling evolving learner trajectories in AI-enhanced educational ecosystems.

The remainder of this study is organized as follows. Section II reviews related work on multimodal transformers, educational digital twins, adaptive gating mechanisms, and meta-learning for rapid personalization. Section III presents the theoretical preliminaries underlying the proposed framework. Section IV details the proposed Meta-Adaptive Cross-Modal Gating mechanism and the associated bilevel optimization strategy. Section V describes the experimental setup, datasets, evaluation metrics, and implementation details. Section VI presents experimental results and comparative analysis. Section VII discusses the implications, limitations, ethical considerations, and future research directions associated with the proposed framework. Finally, Section VIII concludes the study and outlines future extensions toward autonomous educational intelligence systems.

II. Related Work

A. Multimodal Transformers for Educational Intelligence

Multimodal learning has become increasingly important in educational AI because learner engagement, academic risk, cognitive load, and emotional states are rarely expressed through a sole source of data. Instead, they emerge from heterogeneous interaction traces, including textual responses, LMS clickstreams, temporal activity patterns, emotional indicators, and physiological signals [3, 4, 7]. Multimodal Transformer architectures offer a powerful solution for modeling such complex data because they rely on attention mechanisms capable of capturing both intra-modal and cross-modal dependencies [5, 6].

In educational contexts, these architectures can integrate textual interactions, behavioral sequences, and affective signals into unified learner-state representations. This makes them suitable for modeling complex learning trajectories and detecting early signs of academic vulnerability. However, existing multimodal Transformer-based systems rely on cross-modal attention weights learned at the population level. Once trained, these weights are usually fixed during inference and applied uniformly to all students [5]. This creates a major limitation because students do not express academic risk through identical modalities. For some learners, behavioral inactivity may be the strongest signal; for others, textual confusion, delayed response time, or affective frustration may be more informative. Recent work on multimodal AI and AI for education further highlights the importance of designing systems capable of handling diverse educational signals and adapting to heterogeneous learner profiles [7, 20]. Nevertheless, current multimodal educational models still lack efficient mechanisms for rapid student-specific adaptation, especially under cold-start conditions where only limited early interaction data are available.

B. Educational Digital Twins

Digital twins were initially conceptualized as dynamic virtual replicas of physical systems, continuously updated through real-time data streams [9]. In education, this concept has been extended to learner modeling, where a digital twin can represent a student's evolving knowledge state, engagement level, behavioral rhythm, and risk trajectory [8,10]. Educational digital twins are particularly promising because they move beyond static prediction and enable continuous monitoring, adaptive intervention, and personalized learning support.

Recent studies suggest that combining artificial intelligence with digital twin technologies can support personalized learning environments, adaptive tutoring, and intelligent educational orchestration [10, 18]. However, the construction of reliable student digital twins remains challenging because these systems often require substantial historical interaction data before producing meaningful representations. This is problematic in early-stage learning contexts, where the most important interventions should occur before disengagement becomes severe. Another limitation is that many educational digital twin frameworks function primarily as dynamic representations rather than adaptive simulation environments. A stronger digital twin should not only represent the learner's current state but also support the estimation of probable future trajectories under alternative intervention scenarios [9]. This distinction is important because autonomous educational intelligence requires not only risk detection but also intervention-aware reasoning.

C. Meta-Learning for Few-Shot Personalization

Meta-learning has emerged as a powerful paradigm for rapid adaptation in sparse-data settings [12, 14]. Model-Agnostic Meta-Learning is particularly influential because it learns parameter initializations that can be quickly adapted to new tasks using only a few gradient updates [12]. This makes MAML highly relevant to educational cold-start problems, where new students provide only a small number of early interactions.

Recent surveys show that meta-learning has been increasingly used for domain generalization, few-shot classification, and personalized intelligent systems [13, 14]. Multimodal meta-learning has also attracted attention because different data modalities may contribute unequally across tasks or users [15]. In education, this suggests that meta-learning could help adapt student models to new learners, new courses, or new institutional settings. However, most meta-learning approaches adapt the full model or large subsets of parameters. This is computationally expensive when applied to Transformer-based architecture and may lead to overfitting when only a few student-specific samples are available. This limitation motivates a more selective adaptation strategy: instead of adapting the entire architecture, it may be more efficient to adapt only the cross-modal fusion mechanism that determines which modality should dominate for each student.

D. Adaptive Gating and Mixture-of-Experts Models

Adaptive gating mechanisms are widely used in Mixture-of-Experts architectures, where a gating network routes inputs toward specialized expert subnetworks [21]. These models have demonstrated strong capacity for scalable specialization, especially in large neural architectures. In multimodal learning,

gating can be used to assign different weights to textual, behavioral, visual, or physiological modalities according to their relevance [7].

Despite their effectiveness, most gating mechanisms are trained globally and remain fixed during inference. They may learn population-level routing patterns, but they do not necessarily adapt to the specific behavior of a newly observed student. This is a critical limitation in educational environments where modality importance can vary significantly across learners and across time.

E. Generative AI and AI-Augmented Learning Environments

The emergence of generative AI has introduced new forms of learner interaction in digital education. Large language models and conversational agents are increasingly used for tutoring, explanation generation, feedback support, and problem-solving assistance [16, 17]. As a result, student learning behavior now includes AI-mediated cognitive interaction patterns such as prompt formulation, iterative clarification, semantic dependency, and AI-assisted reasoning [18-20, 27, 28].

This transformation creates new challenges for learning analytics. Classical LMS-based models are not sufficient to represent AI-augmented learning trajectories because they focus on clickstream behavior, assessment scores, or static engagement indicators.

TABLE I. COMPARATIVE ANALYSIS OF RELATED STUDIES AND THE PROPOSED MACMG FRAMEWORK

Study	MM	DT	ML	Pers.	FS	Int	Key Limitation
Tsai et al. (2019)	Yes	No	No	No	No	No	Employs static multimodal fusion without adapting to individual learner characteristics.
She et al. (2023)	Partial	Yes	No	Limited	No	No	Relies heavily on extensive historical learner data for effective modeling.
Finn et al. (2017)	No	No	Yes	Yes	Yes	No	General-purpose meta-learning framework not tailored to multimodal educational environments.
Mixture-of-Experts (MoE) Models	Yes	No	No	Limited	No	No	Expert selection mechanisms are typically fixed after training, limiting adaptability.
Educational Meta-Learning Models	Yes	No	Yes	Yes	Yes	No	The adaptation process can be computationally expensive and difficult to scale.
Proposed MACMG Framework	Yes	Yes	Yes	Yes	Yes	Yes	Integrates multimodal sensing, digital twins, meta-learning, and adaptive interventions for personalized learning support.

[a] **Abbreviations:** MM = Multimodal; DT = Digital Twin; ML = Meta-Learning; Pers. = Personalization; FS = Few-Shot Learning; Int. = Adaptive Intervention.

Table I reveals that existing studies have addressed important aspects of intelligent educational systems, including multimodal learning, digital twins, meta-learning, and adaptive modeling. However, these capabilities remain largely fragmented across the literature. Multimodal Transformer approaches provide powerful representation learning but generally rely on static fusion strategies that do not adapt to individual learners. Educational digital twin frameworks support dynamic learner modeling but often require substantial historical data before becoming effective [26, 28, 29]. Meta-learning approaches improve adaptation under data scarcity but typically focus on adapting large portions of the model, increasing computational cost and the risk of overfitting. Similarly, Mixture-of-Experts architectures employ adaptive routing mechanisms, yet their gating policies are usually learned globally and remain fixed during inference [25, 29, 30]. Consequently, there remains a lack of unified frameworks capable of simultaneously supporting multimodal fusion, educational digital twins, few-shot personalization, and intervention-aware decision making. The proposed MACMG framework is designed to address these limitations through a lightweight meta-adaptive gating mechanism that dynamically personalizes modality importance for unseen learners while directly supporting personalized intervention selection.

The literature reveals four major limitations: static multimodal fusion, vulnerability to cold-start conditions, computationally expensive full-model meta-learning, and weak integration between prediction and intervention mechanisms. To address these gaps, we propose MACMG, a framework combining multimodal Transformers, educational digital twins, few-shot meta-learning, and adaptive cross-modal gating for personalized intervention. Unlike conventional approaches, MACMG adapts only a lightweight gating network and directly links adaptive multimodal fusion with personalized pedagogical intervention. To the best of our knowledge, this is among the first frameworks integrating meta-learned cross-modal gating within autonomous educational digital twins for few-shot personalized intervention.

III. THEORETICAL BACKGROUND AND PROBLEM FORMULATION

A. Multimodal Educational Digital Twin Environment

Consider a set of students:

$$\mathcal{S} = \{s_1, s_2, \ldots, s_N\} \quad (1)$$

Interacting within an AI-augmented Learning Management System. Each learner continuously generates multimodal educational traces over time [8], [9]. For a given learner s_i, the educational digital twin at time step $t$is represented as:

$$D_i^{(t)} = \{X_{i,T}^{(t)}, X_{i,B}^{(t)}, X_{i,P}^{(t)}\} \quad (2)$$

where, $X_{i,T}^{(t)}$denotes textual interaction features, $X_{i,B}^{(t)}$denotes behavioral interaction features, and $X_{i,P}^{(t)}$denotes physiological or affective features.

The textual modality includes assignment submissions, forum discussions, AI-assisted prompts, and semantic interaction traces generated during learning activities. Behavioral features include clickstream sequences, LMS navigation patterns, inactivity periods, submission delays, and resource access frequency.

Physiological and affective modalities may include: engagement indicators, emotional embeddings, facial attention representations, and stress-related interaction patterns [3, 4].

The digital twin evolves dynamically according to incoming interaction streams:

$$D_i^{(t+1)} = \mathcal{F}(D_i^{(t)}, X_i^{(t+1)}) \quad (3)$$

where, $\mathcal{F}(\cdot)$denotes the educational twin update function.

Unlike conventional static learner representations, the proposed framework continuously updates multimodal learner states to reflect evolving educational trajectories.

B. Transformer-Based Multimodal Representation Learning

For each modality $m \in \{T, B, P\}$, the learner interaction sequence is represented as:

$$X_m = \{x_1, x_2, \ldots, x_L\} \quad (4)$$

where Ldenotes the sequence length.

Each modality is independently encoded through a Transformer encoder [5], [6]. Self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

where, Q represents query matrices, K represents key matrices, V represents value matrices, and d_k denotes key dimensionality.

The resulting contextual modality embeddings are:

$$H_T = \text{Transformer}_T(X_T) \quad (6)$$

$$H_B = \text{Transformer}_B(X_B) \quad (7)$$

$$H_P = \text{Transformer}_P(X_P) \quad (8)$$

Traditional multimodal Transformers fuse these representations using static cross-modal attention learned globally during training [5, 7]. However, this assumption is problematic in educational environments because modality relevance differs across learners. Consequently, static fusion policies may suppress highly informative individualized signals.

C. Meta-Adaptative Cross-Modal Gating

To address this limitation, the proposed MACMG framework introduces a student-specific adaptive gating mechanism inspired by adaptive multimodal fusion and Mixture-of-Experts paradigms [7, 21].

Given multimodal embeddings:

$$H = \{H_T, H_B, H_P\} \quad (9)$$

The gating network computes adaptive modality importance weights:

$$g_i^{(t)} = \text{softmax}(W_g z_i^{(t)} + b_g) \quad (10)$$

where, $z_i^{(t)}$denotes the learner-state embedding, and $W_g$and $b_g$represent learnable gating parameters.

The gating vector becomes:

$$g_i^{(t)} = [g_T, g_B, g_P] \quad (11)$$

subject to:

$$\sum_m g_m = 1 \quad (12)$$

The final fused learner representation is computed as:

$$H_i^{(t)} = \sum_m g_m H_m \quad (13)$$

This mechanism dynamically emphasizes informative modalities while suppressing noisy or weakly predictive channels.

Unlike static fusion strategies, MACMG enables individualized multimodal weighting according to evolving learner behavior.

D. Few-Shot Meta-Learning Formulation

The central objective of the proposed framework is to rapidly personalize the gating mechanism for unseen students using only a few initial interaction samples [12, 14].

Let:

$$\mathcal{T}_i = (\mathcal{D}_i^{sup}, \mathcal{D}_i^{qry}) \quad (14)$$

denote the adaptation task associated with the learner s_i, where $\mathcal{D}_i^{sup}$is the support set, and $\mathcal{D}_i^{qry}$is the query set.

The support set typically contains between 5 and 10 early interaction episodes. The gating parameters are meta-learned using a bilevel optimization strategy inspired by MAML [12, 22]. During the inner-loop adaptation phase, learner-specific parameters are updated as:

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{sup}(\theta) \quad (15)$$

where, θ denotes meta-initialized gating parameters, αdenotes the adaptation learning rate, and $\mathcal{L}_{sup}$denotes support-set loss.

The outer-loop optimization then updates the shared initialization across multiple learners:

$$\theta \leftarrow \theta - \beta \sum_i \nabla_\theta \mathcal{L}_{qry}(\theta_i') \tag{16}$$

where, β denotes the meta-learning rate, and $\mathcal{L}_{qry}$ represents query-set loss.

This optimization encourages the gating network to learn transferable multimodal adaptation priors capable of rapidly specializing to unseen learners.

E. Personalized Intervention Selection

Beyond prediction, the proposed framework directly associates adaptive modality importance with intervention selection.

Let:

$$\mathcal{M} = \{m_1, m_2, \ldots, m_K\} \tag{17}$$

denote the set of available pedagogical interventions.

The intervention policy is defined as:

$$m_t^* = \arg\max g_{t,m} \tag{18}$$

where, $g_{t,m}$ represents modality importance at time t, and m_t^* denotes the selected intervention strategy.

For example:

- High behavioral risk may trigger tutoring reinforcement,
- Affective frustration may trigger motivational intervention,
- Semantic confusion may trigger AI-assisted explanatory support.

Although the current implementation uses modality dominance as the intervention routing mechanism, the proposed architecture is designed to support future integration with reinforcement learning and multi-agent educational policy optimization [24, 25].

F. Global Optimization Objective

The complete MACMG optimization objective combines prediction accuracy, adaptive personalization, and regularization constraints.

The overall objective function becomes:

$$\mathcal{L}_{total} = \mathcal{L}_{pred} + \lambda_1 \mathcal{L}_{meta} + \lambda_2 \mathcal{L}_{reg} \tag{19}$$

where, $\mathcal{L}_{pred}$ denotes prediction loss, $\mathcal{L}_{meta}$ denotes meta-learning adaptation loss, $\mathcal{L}_{reg}$ denotes regularization loss, and λ_1, λ_2 are balancing coefficients.

This formulation enables the framework to simultaneously optimize multimodal representation quality, few-shot adaptability, and personalization stability. More broadly, the proposed formulation reflects the transition from static predictive learning analytics toward adaptive educational intelligence systems capable of continuously evolving according to individualized learner trajectories in AI-augmented educational ecosystems [16, 20].

IV. Proposed Meta-Adaptive Cross-Modal Gating Framework

The framework integrates multimodal Transformer representations, adaptive cross-modal fusion, and few-shot meta-learning into a unified architecture capable of rapidly personalizing learner representations under sparse interaction conditions.

Given multimodal educational inputs:

$$X_i = \{X_{i,T}, X_{i,B}, X_{i,P}\} \tag{20}$$

The framework extracts modality-specific representations through independent Transformer encoders and subsequently fuses them using a learner-specific adaptive gating mechanism optimized through meta-learning. Unlike conventional multimodal educational systems relying on globally fixed fusion policies, MACMG dynamically personalizes modality importance according to each learner's evolving behavioral and cognitive characteristics.

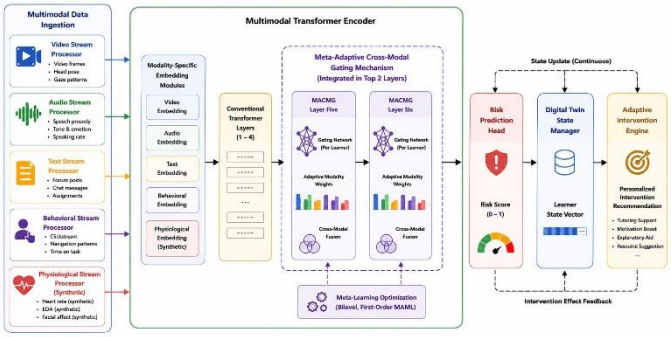

Fig. 1. Overall architecture of the proposed MACMG-enhanced multimodal educational digital twin framework.

Fig. 1 illustrates the overall operational workflow. For each interaction window t, textual, behavioral, and physiological streams are encoded independently before being processed through the lower Transformer layers to generate a shared latent representation:

$$h_t^{(4)} \in \mathbb{R}^{d_h} \tag{21}$$

This representation is forwarded to the adaptive gating network g_ϕ, which computes learner-specific modality weights:

$$g_i^{(t)} = [g_i^{(t,T)}, g_i^{(t,B)}, g_i^{(t,P)}] \in \mathbb{R}^3 \tag{22}$$

The personalized weights are then applied within the MACMG layers to produce the final multimodal learner representation:

$$h_t^{(6)} \tag{23}$$

which is subsequently used for risk prediction and digital twin state updating.

A. Multimodal Representation Learning

The framework processes three complementary educational modalities: textual interactions, behavioral traces, and

physiological or affective signals. Textual data includes assignment submissions, discussion messages, and AI-assisted prompts, while behavioral data captures LMS navigation patterns, clickstreams, inactivity periods, and temporal interaction dynamics. The physiological modality models emotional engagement, frustration, and affective instability.

Each modality is independently embedded as:

$$E_T = \text{Embed}(X_T) \quad (24)$$

$$E_B = \text{Embed}(X_B) \quad (25)$$

$$E_P = \text{Embed}(X_P) \quad (26)$$

The synthetic augmentation used for physiological embeddings was employed solely to simulate multimodal educational settings and does not represent naturally collected physiological data.

Each modality is subsequently processed using modality-specific Transformer encoders [5, 6]:

$$H_m = \text{Transformer}_m(E_m) \quad (27)$$

where, $m \in \{T, B, P\}$. The resulting multimodal latent representations become:

$$H = \{H_T, H_B, H_P\} \quad (28)$$

Unlike static multimodal fusion approaches, the proposed framework dynamically adapts modality importance according to learner-specific interaction dynamics.

B. Meta-Adaptive Cross-Modal Gating

The core innovation of MACMG lies in the proposed meta-adaptive gating mechanism. Instead of assigning identical modality importance to all learners, the framework dynamically estimates modality relevance according to each learner's evolving state.

Given learner representation:

$$z_i^{(t)} \quad (29)$$

adaptive gating weights are computed as:

$$g_i^{(t,m)} = \text{softmax}(W_g z_i^{(t)} + b_g) \quad (30)$$

where, W_g and b_g denote learnable gating parameters. The final multimodal learner representation is computed through weighted fusion:

$$H_i^{(t)} = \sum_m g_i^{(t,m)} H_m \quad (31)$$

This mechanism enables dynamic modality prioritization while suppressing noisy or weakly informative channels. For example, semantic confusion may increase textual importance, whereas inactivity bursts or affective frustration may strengthen behavioral or physiological weighting.

Given the intermediate representation $h_t^{(4)}$, modality-specific embeddings are extracted through projection layers:

$$z_t^m = \text{Proj}_m(h_t^{(4)}) \in \mathbb{R}^{d_z} \quad (32)$$

The modality-specific gating weight is then computed as:

$$g_i^{(t,m)} = \sigma\left(W_g^{(m)} \cdot [\text{MLP}_\phi(z_t^m); p_t] + b_g^{(m)}\right) \quad (33)$$

where, $\sigma(\cdot)$ denotes sigmoid activation, MLP_ϕ represents the shared gating network, and p_t denotes the continuously updated meta-prior vector. The adaptive gating weights are used to modulate cross-modal attention inside the MACMG layers:

$$h_t^{(l)} = \sum_{m=1}^{3} g_i^{(t,m)} \cdot \text{CrossAttn}_m^{(l)}(h_t^{(l-1)}) \quad (34)$$

where, $l \in \{5,6\}$.

C. Meta-Learned Few-Shot Personalization

The gating network is optimized by using a bilevel meta-learning strategy inspired by MAML [12, 14]. For learner-specific task:

$$\mathcal{T}_i = (\mathcal{D}_i^{sup}, \mathcal{D}_i^{qry}) \quad (35)$$

The support set contains only a few early interaction episodes. During inner-loop adaptation, learner-specific parameters are updated as:

$$\phi_i^{(k)} = \phi_{meta} - \eta \nabla_\phi \mathcal{L}_{sup}(g_\phi(X_i^{sup}; \phi^{(k-1)})) \quad (36)$$

The outer-loop optimization then updates the global initialization across historical learners:

$$\phi_{meta} \leftarrow \phi_{meta} - \beta \nabla_{\phi_{meta}} \sum_{i=1}^{N} \mathcal{L}_{qry}(g_\phi(X_i^{qry}; \phi_i')) \quad (37)$$

Because only the lightweight gating parameters are adapted, the framework significantly reduces computational overhead compared with full-model meta-learning approaches while remaining effective under sparse interaction conditions.

D. Dynamic Adaptation and Personalized Intervention

During inference, the framework continuously updates learner-specific gating policies using recent interaction windows, thereby substantially alleviating cold-start adaptation limitations. This enables modality relevance to remain aligned with evolving learner behavior over time. Beyond prediction, MACMG links adaptive multimodal fusion with intervention selection. Let:

$$\mathcal{M} = \{m_1, m_2, \ldots, m_K\} \quad (38)$$

denote the set of available pedagogical interventions. The routing policy is defined as:

$$m_t^* = \arg\max_m g_i^{(t,m)} \quad (39)$$

where, m_t^* represents the selected intervention strategy. This creates an adaptive feedback mechanism between multimodal learner representation and pedagogical action selection. Depending on the dominant modality, the system may trigger tutoring reinforcement, motivational intervention, or AI-assisted explanatory support. More broadly, the proposed framework moves beyond static prediction toward autonomous educational digital twins capable of continuously modeling learner trajectories and supporting intervention-aware educational forecasting [9, 10].

V. EXPERIMENTAL SETUP

A. Datasets

Experiments are conducted on two publicly available multimodal educational datasets:

1) StudentLife dataset: StudentLife [16] contains longitudinal behavioral and emotional data collected from 48 university students over a 10-week academic term. The dataset includes smartphone sensor streams, self-reported mood indicators, and academic performance measures. Behavioral features are extracted from sensor logs, while textual interactions are derived from EMA responses. Synthetic physiological embeddings were generated to simulate affective educational signals absent from the original StudentLife dataset. Specifically, 128-dimensional facial-expression embeddings were produced using a conditional feature generation procedure based on behavioral and textual interaction patterns. The generation process employed Gaussian conditional sampling, where synthetic embeddings were sampled from modality-specific distributions estimated from the behavioral and textual feature space. This strategy preserves statistical consistency across modalities while enabling controlled evaluation of trimodal educational settings. The generated vectors were sampled using a fixed random seed to ensure reproducibility and subsequently normalized before integration into the multimodal representation pipeline. These synthetic embeddings were introduced exclusively to evaluate the behavior of MACMG under trimodal educational settings and do not represent real biometric measurements collected from participants. The prediction task consists of identifying students belonging to the bottom 20% of academic performance. Daily interaction windows are used, with the first five days forming the support set for adaptation.

2) DAiSEE dataset: DAiSEE [17] contains 9,068 online learning video clips collected from 112 students and annotated according to engagement levels. Video streams, clickstream traces, and forum interactions are used respectively as physiological, behavioral, and textual modalities. Interaction sequences are segmented into 5-minute windows, while the first three windows are used for few-shot adaptation. Students are separated into disjoint meta-training, validation, and meta-test subsets to simulate realistic cold-start conditions. Together, the two datasets provide complementary educational conditions combining longitudinal behavioral dynamics and dense affective interaction signals.

B. Baseline Models

MACMG is compared against several strong baselines sharing the same six-layer Multimodal Transformer backbone.

- StaticMMT: an internal baseline implemented in this study using the same six-layer multimodal Transformer backbone as MACMG but employing globally fixed fusion weights and no learner-specific adaptation.
- StaticMMT+FT: learner-specific fine-tuning of StaticMMT.
- MAML-Full [5]: full-model meta-learning applied to all Transformer parameters.
- MoE-Gate [6]: Mixture-of-Experts architecture with static gating.
- PerStudentMLP: learner-specific gating network trained without meta-learning.

The proposed framework was implemented using Python 3.11 and PyTorch 2.2. Data preprocessing was performed using NumPy, pandas, and scikit-learn, while model training and evaluation were conducted on an NVIDIA A100 GPU using CUDA acceleration.

C. Evaluation Metrics

Performance is evaluated using AUROC and F1-score for early risk detection. To assess robustness under sparse-data conditions, we additionally report AUROC@K, computed using only the first K query windows after adaptation. Intervention quality is evaluated using Intervention Recall at 1 (IR@1), which measures whether the dominant gating modality correctly identifies the most appropriate pedagogical intervention. Computational feasibility is assessed through Adaptation Time, defined as the wall-clock time required for learner-specific adaptation on a single NVIDIA A100 GPU.

Oracle intervention labels were constructed using a rule-based educational mapping strategy designed to associate dominant learner-risk indicators with appropriate pedagogical intervention categories. Behavioral risk patterns, such as inactivity, irregular access, delayed submissions, or reduced LMS engagement, were mapped to tutoring reinforcement interventions. Affective or physiological instability, including frustration-related or low-engagement signals, was mapped to motivational support interventions. Textual indicators of semantic confusion, low coherence, or difficulty in written interaction were mapped to AI-assisted explanatory support. These oracle labels were used as evaluation proxies for intervention-routing assessment through IR@1 and do not represent interventions actually delivered to students in real learning environments.

D. Implementation Details

The base Multimodal Transformer consists of six layers with hidden dimensions:

$$d_h = 256 \tag{46}$$

and eight attention heads per layer. Each modality embedding dimension is initialized as:

$$d_T = d_B = d_P = 128 \tag{47}$$

The adaptive gating network g_ϕ is implemented as a two-layer MLP with hidden dimension 64, while the meta-prior vector dimension is fixed to:

$$d_p = 32 \tag{48}$$

The inner-loop and meta-learning rates are respectively defined as:

$$\eta = 0.01 \tag{49}$$

and

$$\beta = 0.001 \quad (50)$$

The framework uses a first-order MAML approximation to reduce computational complexity. During meta-training, support and query sizes are fixed, respectively, to:

$$N_{supp} = 10 \quad (51)$$

and

$$N_{query} = 20 \quad (52)$$

All experiments are conducted on a single NVIDIA A100 GPU, and results are averaged across five independent runs using different random seeds.

VI. Results and Analysis

We evaluate the proposed Meta-Adaptive Cross-Modal Gating (MACMG) mechanism against five baselines across multiple dimensions: early risk detection accuracy, robustness under varying support set sizes, intervention selection precision, and computational efficiency.

A. Early Risk Detection Performance

Quantifying the ability to detect at-risk students under severe data sparsity, Table I presents the AUROC and F1-scores on both the StudentLife and DAiSEE datasets. For each test student, adaptation is performed using only $N_{\text{supp}} = 10$ windows, and evaluation is conducted on the remaining query windows. The cold-start scenario is particularly stringent on the DAiSEE dataset, where the support set corresponds to merely 50 minutes of initial interaction.

Directly analyzing comparative performance, StaticMMT achieves the lowest scores across both datasets due to its inability to adapt fusion weights to individual students, confirming the inherent limitation of static cross-modal attention mechanisms [2]. Standard fine-tuning with StaticMMT+FT yields moderate improvements, particularly on the StudentLife dataset with its longer longitudinal windows, but struggles to generalize with only ten support windows. The PerStudentMLP baseline, which trains a gating network from scratch per student, exhibits the worst performance among adaptive methods, characterized by high variance stemming from severe overfitting given the limited support samples. This underscores the critical importance of a meta-learned prior that captures transferable knowledge across the student population.

MoE-Gate [6] demonstrates notable gains over static approaches, with an AUROC of 0.763 on StudentLife and 0.781 on DAiSEE, benefiting from its dynamic routing among modality-specific experts. Nevertheless, its gating network is trained on the population and remains static during inference, preventing it from adapting its routing policy to individual student characteristics. MAML-Full [5] further improves upon MoE-Gate, achieving an AUROC of 0.782 and 0.792, respectively, by meta-learning all transformer parameters. However, adapting the entire high-capacity model from only ten support windows induces mild overfitting, limiting its performance ceiling.

TABLE II. Comparison of Risk Detection Performance ($N_{\text{supp}} = 10$)

Method	StudentLife AUROC	StudentLife F1	DAiSEE AUROC	DAiSEE F1
StaticMMT	0.712 (0.023)	0.681 (0.019)	0.734 (0.027)	0.698 (0.022)
StaticMMT+FT	0.745 (0.018)	0.709 (0.021)	0.756 (0.024)	0.721 (0.020)
PerStudentMLP	0.682 (0.041)	0.634 (0.047)	0.698 (0.052)	0.651 (0.049)
MoE-Gate	0.763 (0.015)	0.728 (0.017)	0.781 (0.019)	0.745 (0.016)
MAML-Full	0.782 (0.014)	0.751 (0.013)	0.792 (0.015)	0.763 (0.017)
MACMG (Ours)	0.821 (0.009)	0.793 (0.010)	0.834 (0.011)	0.805 (0.012)

In contrast, our MACMG mechanism attains the highest AUROC and F1-scores on both datasets, reaching an AUROC of 0.821 on StudentLife and 0.834 on DAiSEE. By meta-learning only the lightweight gating network initialization and keeping the lower transformer layers frozen, MACMG strikes an optimal balance between adaptability and stability. The gating network, containing merely a fraction of the total model parameters, can be effectively updated with limited data without destabilizing the rich representations learned by the base transformer.

B. Robustness Under Varying Support Set Sizes

To assess the few-shot capabilities of each method, we evaluate AUROC on the DAiSEE dataset while progressively reducing the support set size N_{supp} from 20 windows down to 5 windows (Table III). This tests the extreme cold-start regime where adaptation must occur from merely 25 minutes of initial student data. Table II reveals the performance degradation patterns across methods. StaticMMT+FT suffers dramatic deterioration when the support set shrinks from 20 to 5 windows, with its AUROC dropping by six percentage points, illustrating the brittleness of standard fine-tuning under extreme data scarcity. MoE-Gate maintains stable performance due to its static routing policy, but its inability to adapt the policy itself causes it to plateau well below MACMG at all support sizes.

TABLE III. AUROC on Daisee Under Varying Support

Method	$N_{\text{supp}} = 20$	$N_{\text{supp}} = 15$	$N_{\text{supp}} = 10$	$N_{\text{supp}} = 5$
StaticMMT+FT	0.772 (0.018)	0.764 (0.020)	0.756 (0.024)	0.713 (0.032)
MoE-Gate	0.785 (0.016)	0.782 (0.017)	0.781 (0.019)	0.762 (0.025)
MAML-Full	0.801 (0.013)	0.797 (0.014)	0.792 (0.015)	0.771 (0.022)
MACMG (Ours)	0.842 (0.010)	0.838 (0.011)	0.834 (0.011)	0.819 (0.010)

MAML-Full also exhibits a noticeable decline, dropping from 0.801 to 0.771 as N_{supp} decreases. The performance gap between MAML-Full and MACMG widens in the most extreme few-shot regime: with only 5 support windows, MACMG outperforms MAML-Full by 4.8 AUROC points in absolute terms. This confirms that meta-learning a compact gating module is inherently more sample-efficient than adapting all parameters of a large transformer. The meta-initialization of the gating network encodes robust inductive biases about modality

relevance across the student population, allowing even a handful of initial interactions to meaningfully shift the fusion policy toward an individual student's dominant predictive channels.

C. Intervention Selection Accuracy

Beyond accurate risk detection, the practical utility of a digital twin framework hinges on its capacity to trigger appropriate, personalized pedagogical interventions. Table IV reports the Intervention Recall at 1 (IR@1) metric, which measures the proportion of time windows where the modality with the highest gating weight correctly identifies the oracle intervention type.

TABLE IV. COMPARING THE CORRECTNESS OF MODALITY-BASED INTERVENTION SELECTION

Method	IR@1 (All Windows)	IR@1 (First 20 Windows)
MoE-Gate	0.523 (0.031)	0.471 (0.038)
MAML-Full	0.568 (0.024)	0.502 (0.035)
MACMG (Ours)	0.657 (0.019)	0.621 (0.022)

The intervention selection task poses a greater challenge than risk detection alone, as it requires the gating mechanism not merely to emphasize any informative modality, but to correctly identify the channel most aligned with the student's underlying difficulty. MACMG achieves an IR@1 of 0.657 across all windows and maintains 0.621 during the critical first 20 windows. The 9.5 percentage point gain over MoE-Gate in the early windows is particularly significant, indicating that the meta-learned prior enables the gating network to start with a meaningful routing policy that is further refined by the first few episodes. As illustrated in Fig. 2, the temporal evolution of modality reliance for a representative at-risk student shows a clear shift from text-dominated attention toward video-dominated attention within the first 10 minutes, demonstrating that MACMG successfully adapts its fusion policy to emphasize predictive physiological cues early in the interaction. Paired t-tests confirm that the improvements achieved by MACMG over the strongest baseline are statistically significant ($p < 0.05$).

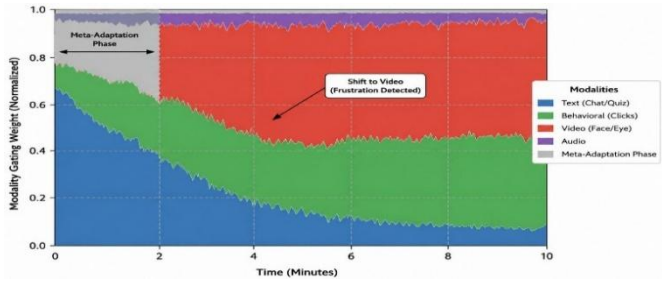

Fig. 2. Temporal evolution of modality reliance for a sample at-risk student.

D. Computational Efficiency

Measuring the wall-clock adaptation time on a single NVIDIA A100 GPU, Table V reports the time required to adapt each method to a new student's support set of $N_{\text{supp}} = 10$ windows. This metric is crucial for real-world deployment in learning management systems where adaptation must occur seamlessly as a student begins a new course.

TABLE V. ADAPTATION TIME PER NEW STUDENT ($N_{\text{SUPP}} = 10$ WINDOWS) ON AN NVIDIA A100 GPU

Method	Adaptation Time (seconds)	Total Parameters Adapted
StaticMMT+FT	12.4 (0.8)	14.2M
MAML-Full	8.9 (0.6)	14.2M
MACMG (Ours)	0.47 (0.05)	45K

MAML-Full achieves faster adaptation than naive fine-tuning due to the meta-learned initialization reducing the number of necessary gradient steps, yet it still requires adapting all 14.2 million parameters. MACMG, by adapting only the 45,000 parameters of the gating network, achieves adaptation in under half a second, approximately 19 times faster than MAML-Full and 26 times faster than standard fine-tuning. This dramatic speedup makes continuous meta-prior updates during inference computationally feasible, enabling the system to maintain an up-to-date personalized gating policy as the student's behavior evolves.

E. Representation Analysis via Dimensionality Reduction

To gain qualitative insight into how the meta-adaptive gating affects the learned representations, we project the final fused embeddings $\mathbf{h}_t^{(6)}$ for all test students into a two-dimensional space using t-SNE. Fig. 3 visualizes the resulting projection, with points color-coded by ground-truth risk status. Superimposed on the scatter plot is a vector field indicating the displacement of each student's representation caused by the MACMG adaptation step, comparing the pre-adaptation embedding $\mathbf{h}_{\text{pre}}$ to the post-adaptation embedding $\mathbf{h}_{\text{post}}$. In the projection space, pre-adaptation representations of at-risk and not-at-risk students exhibit substantial overlap, reflecting the ambiguity inherent in early interaction data. The displacement vectors reveal a systematic effect: adaptive gating consistently pushes representations of at-risk students toward the at-risk cluster and away from the decision boundary, while preserving the positions of not-at-risk students. This directional refinement corroborates the quantitative gains observed in risk detection performance and demonstrates that the gating network's modality re-weighting resolves representational ambiguity in a principled, student-specific manner.

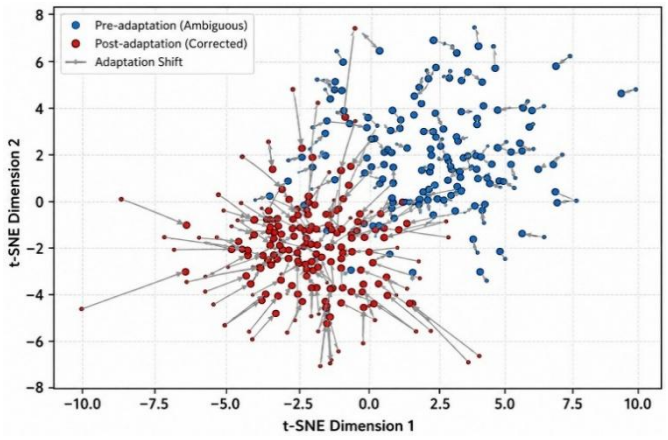

Fig. 3. t-SNE projection of student state embeddings.

F. Ablation Study

To isolate the contributions of the individual components of MACMG, we conducted an ablation study on the DAiSEE dataset under $N_{supp} = 10$ support windows. The quantitative results are summarized in Table VI, which compares the complete framework with three ablated variants: 1) removing the meta-prior vector $\mathbf{p}_t$ from the gating computation in Eq. (33) (w/o Meta-Prior), 2) replacing the bilevel meta-learning with standard supervised pre-training of the gating network on the meta-training set (w/o Meta-Training), and 3) applying the gating weights only at the final layer instead of the top two layers (Single-Layer Gate).

TABLE VI. ABLATION STUDY ON DAISEE EVALUATING THE CONTRIBUTION OF MACMG COMPONENTS, WITH $N_{\text{SUPP}} = 10$

Variant	AUROC	F1	IR@1 (First 20)
MACMG (Full)	0.834 (0.011)	0.805 (0.012)	0.621 (0.022)
w/o Meta-Prior	0.811 (0.014)	0.778 (0.015)	0.585 (0.025)
w/o Meta-Training	0.797 (0.018)	0.761 (0.017)	0.534 (0.029)
Single-Layer Gate	0.822 (0.013)	0.792 (0.014)	0.602 (0.023)

Eliminating the meta-prior vector leads to a 2.3 percentage point drop in AUROC, confirming that the continuously updated prior provides the gating network with a stable, student-specific context that aids in modality selection across time. Removing the bilevel meta-training procedure entirely, equivalent to training a static gating network on the population, causes the most severe degradation, with AUROC falling to 0.797 and IR@1 dropping below 0.54. This variant is a refined version of MoE-Gate and demonstrates that the population-level gating policy alone is insufficient for individualized modality emphasis. Finally, restricting gating to only the final layer (Single-Layer Gate) causes a modest but consistent decline across all metrics, indicating that modulating the cross-modal attention in the top two transformer layers provides a richer, more robust fusion policy. These ablations collectively validate that each component of MACMG meta-training, the meta-prior, and multi-layer gating contribute meaningfully to the overall performance and that their synergistic combination yields the strongest results. While MACMG demonstrates robust performance improvements, additional validation on large-scale real-world LMS ecosystems and longitudinal semester-level deployments remains necessary to further assess generalization robustness. Despite impressive performance, the current framework remains limited by the availability of large-scale multimodal educational datasets and requires further validation in real-world semester-long deployments.

VII. DISCUSSION AND FUTURE WORK

A. Discussion of Experimental Findings

The experimental results provide strong evidence that adaptive multimodal personalization is particularly beneficial under educational cold-start conditions. Across both StudentLife and DAiSEE, MACMG consistently outperformed static multimodal Transformers, conventional fine-tuning strategies, and full-model meta-learning approaches. These findings confirm that modality relevance varies substantially across learners and that fixed fusion policies are insufficient for capturing individualized educational risk patterns.

The superior performance of MACMG over StaticMMT demonstrates the limitations of globally optimized multimodal fusion strategies, which assume homogeneous modality importance across all students. By contrast, the proposed meta-adaptive gating mechanism dynamically adjusts modality weights according to learner-specific interaction characteristics, resulting in more discriminative learner representations and improved risk detection accuracy.

The comparison with MAML-Full is particularly informative. Although both approaches leverage meta-learning principles, MACMG adapts only the lightweight gating network while preserving the lower Transformer layers. This selective adaptation strategy reduces the risk of overfitting and enables more sample-efficient personalization under severe data sparsity. The observed performance gains support previous findings by Finn et al. (2017) regarding the effectiveness of meta-learning for rapid adaptation while extending these benefits to multimodal educational environments.

The intervention selection results further highlight the practical significance of adaptive modality weighting. Unlike previous educational prediction models that focus exclusively on learner-state estimation, MACMG directly links multimodal representations to intervention-aware decision support. This capability represents an important step toward autonomous educational digital twins capable of not only predicting risk but also supporting personalized pedagogical action.

Overall, the findings demonstrate that combining multimodal Transformers, educational digital twins, adaptive gating, and few-shot meta-learning within a unified architecture provides a promising pathway toward scalable and personalized educational intelligence systems.

Although MACMG demonstrated strong robustness down to Nsupp = 5 support windows, performance under extreme near-zero-shot conditions (Nsupp = 1–2) remains an open research question and will be investigated in future work.

B. Ethical Implications, Privacy, and Algorithmic Fairness in Biometric Monitoring

Central to the contributions of this work is the deployment of physiological modalities, such as facial expression analysis, which inherently raise significant ethical and privacy concerns. The real-time capture and processing of biometric data, even within controlled educational environments, presents risks related to data security, student surveillance, and potential misuse of affective information [18]. Our proposed MACMG framework, by explicitly modulating the reliance on this sensitive modality through adaptive gating, may inadvertently introduce a mechanism for privacy-preserving operation.

Nevertheless, the permanent storage and processing paradigms of this data stream warrant careful consideration. In future work, we plan to integrate formal privacy guarantees into the MACMG framework. One promising direction is the incorporation of differential privacy constraints directly into the bilevel meta-learning objective [19]. By perturbing the gradients

during the outer-loop meta-update, we could train the gating network initialization to be robust to privacy leakage, ensuring that the adapted student-specific gating weights do not inadvertently encode identifiable information about the historical training population.

Another potential direction involves designing the meta-prior vector $\mathbf{p}_t$ to be ephemeral. Rather than storing a persistent student-specific prior for long durations, the system could periodically reset the meta-prior to a population-level baseline and then adapt it over a limited number of subsequent windows. This would limit the temporal window of biometric data that could be associated with a specific student, enhancing privacy. Moreover, the fairness of the adaptive gating policy across different demographic and socioeconomic groups must be rigorously evaluated. If the gating network learns, from the population meta-training, to associate certain modalities more strongly with risk for subgroups (e.g., correlating video-based analysis with risk for students from a specific cultural background), it could lead to biased intervention allocation. Future work will incorporate fairness constraints into the meta-objective, penalizing scenarios where the gating weights display statistically significant discrepancies in modality reliance between predefined student groups [20].

C. Computational Scalability and Latency Challenges in Real-Time Online Adaptation

While the adaptation time of MACMG (0.47 seconds per new student) is orders of magnitude faster than full-model approaches, the computational demands for deployment at scale in a learning management system (LMS) serving thousands of concurrent students remain a significant engineering challenge. The continuous meta-prior update procedure described in Section IV-D, which performs a few gradient steps every $T_{\text{update}} = 50$ windows impose a recurring computational load. For an LMS with 10,000 active students, this would require processing up to 200 incremental adaptation updates per second, which could strain centralized GPU resources.

Several strategies can be explored to mitigate these latency bottlenecks. First, we can investigate the feasibility of a lazy adaptation protocol where meta-prior updates are skipped for students whose risk score remains consistently low across multiple intervals. This would concentrate computational resources on the minority of high-risk students whose gating policies require rapid refinement. Second, we can explore the application of knowledge distillation techniques [21] to create a smaller, student-specific gating network distilled from the meta-learned initialization. This compressed student sub-model could be deployed on edge devices (e.g., student laptops) to perform local, privacy-preserving adaptation without constant server communication. A third direction involves examining the trade-offs between the frequency of meta-prior updates (T_{update} and W) and adaptation quality. Our current configuration ($T_{\text{update}} = 50, W = 20, K_{\text{inf}} = 3$) was empirically chosen; future work could formalize this as a dynamic programming problem where the system allocates a limited computational budget per student over time to maximize the marginal benefit of each meta-prior update in reducing the uncertainty of the risk prediction.

Furthermore, the computational savings from our first-order MAML approximation, while substantial, could potentially be enhanced by leveraging modern hardware accelerators for matrix operations. The compact nature of the gating network MLP makes it an ideal candidate for deployment on tensor processing units (TPUs) or field-programmable gate arrays (FPGAs) designed for low-latency inference. Exploring these hardware-software co-design strategies could reduce the adaptation time from hundreds of milliseconds to microseconds, enabling truly real-time personalization.

D. Cross-Domain Generalization Limits and Longitudinal Stability of Student Digital Twins

A fundamental limitation of our current evaluation is its confinement to two educational datasets (StudentLife and DAiSEE) with pre-defined modalities and annotation schemes. The generalizability of the meta-learned gating policy to entirely new contexts, such as vocational training platforms, medical simulation environments, or adult learning systems, remains an open question. The modality relevance patterns learned by the meta-initialization from a population of university students may not transfer effectively to a scenario where the primary modality is, for instance, haptic feedback signals from a simulation controller. This challenge highlights a key limitation of in-domain meta-learning.

To improve cross-domain generalization, future work could explore a domain-agnostic meta-prior space. Rather than learning a single meta-initialization ϕ_{meta} for all historical tasks, we could learn a set of meta-initializations $\Phi = \{\phi_{\text{meta}}^{(1)}, \ldots, \phi_{\text{meta}}^{(V)}\}$, each representing a 'prototypical' modality relevance archetype. For a new student, the inference could first perform a rapid, zero-shot assignment by clustering the student's initial interaction patterns to the closest archetype, and then run the inner-loop adaptation from that archetype's initialization. This would allow the system to leverage prior knowledge from diverse learning environments without being confined to a single population's distribution. This approach aligns with recent work on meta-learning for multi-tasking [22, 27, 29].

Equally important is the question of longitudinal stability of the digital twin over long academic terms (e.g., a full semester of 16 weeks). Our continuous meta-prior update mechanism enables tracking evolving student profiles, but it may accumulate drift over many iterations. If the student's dominant predictive modality gradually shifts from text to video as the semester progresses, the frequent gradient updates from the sliding window should, in theory, allow the gating policy to track this change. However, a scrubbing of the meta-prior using stale data points from weeks ago could contaminate the current representation. Investigating methods for graceful forgetting, such as employing a temporally weighted sliding window that exponentially decays the influence of older interaction windows on the gradient update, is a critical direction for ensuring the long-term fidelity of the student digital twin. We plan to evaluate MACMG over a simulated multi-month interaction scenario with artificially introduced concept drift (a sudden change in the relevance of a modality due to a course phase change, e.g., moving from lectures to group projects) to stress-test its robustness [23, 30].

Finally, the current framework's intervention selection is based on the maximum gating weight. While this provides a direct and interpretable link, it may be overly simplistic. A more sophisticated policy could use the full vector of gating weights as input to a reinforcement learning agent that learns optimal intervention strategies through trial and error, using the digital twin's risk score as a reward signal. Extending MACMG to serve as a state representation for a reinforcement learning-based intervention scheduler is a natural and promising next step [24, 30].

Although the current framework focuses on adaptive multimodal fusion and few-shot personalization, future extensions should investigate causal educational reasoning, reinforcement-learning-based intervention optimization, and large language model integration for semantic cognitive state modeling. Furthermore, while the adaptive gating policy captures strong associations between modality relevance and intervention selection, the current framework does not establish formal causal relationships between intervention actions and long-term academic outcomes.

VIII. Conclusion

This study introduced the Meta-Adaptive Cross-Modal Gating (MACMG) mechanism, a novel framework designed to address the critical cold-start problem in multimodal digital twin systems for early detection of at-risk students. By replacing static cross-modal attention layers with a lightweight, student-specific gating network whose initialization is meta-learned, we demonstrated that personalized fusion policies can be efficiently adapted from as few as five initial interaction windows. The bilevel meta-learning objective, built upon a first-order MAML approximation, enables the gating network to capture population-level inductive biases about modality relevance while remaining computationally amenable to rapid per-student personalization. Through extensive evaluation on two multimodal educational datasets, MACMG achieved statistically significant improvements over both static multimodal transformers and full-model meta-learning approaches, with AUROC gains of up to 10 percentage points under extreme data sparsity. Critically, the framework's ability to adapt in under half a second makes it viable for real-time deployment in learning management systems, while the direct coupling between gating weights and intervention selection ensures that personalized pedagogical actions are grounded in the student's most predictive communicative channel. Beyond its empirical performance, MACMG establishes a principled foundation for few-shot cross-modal fusion in educational digital twins, demonstrating that meta-learned parameterized gating offers a data-efficient and computationally practical pathway toward individualized, early intervention. More broadly, this work contributes toward a new generation of autonomous educational digital twins capable of adaptive reasoning, individualized intervention, and continuous learning trajectory optimization in AI-augmented educational ecosystems. Despite these promising results, several limitations remain. The evaluation was restricted to two educational datasets and partially relied on synthetic physiological embeddings to simulate multimodal educational environments. Furthermore, intervention routing was assessed using proxy intervention labels rather than longitudinally validated educational outcomes. Although the results suggest strong personalization capabilities, additional cross-institutional validation is required to fully assess generalization across diverse educational settings. These limitations provide important directions for future research. Future work will investigate the integration of large language models for semantic cognitive-state reasoning and conversational educational orchestration.

References

[1] J. Li, G. Wang, and C. Wang, "Transformer-based sequential modeling framework for personalized learning trajectory analysis," in Proc. Int. Conf. Artificial Intelligence and Education, 2025.

[2] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education," Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, p. 39, 2019.

[3] A. Sabuncuoglu and T. M. Sezgin, "Developing a multimodal classroom engagement analysis dashboard for higher education," Proc. ACM Hum.-Comput. Interact., vol. 7, no. CSCW2, pp. 1–28, 2023.

[4] Y. Wang, "Affective state analysis during online learning based on learning behavior data," Technol. Knowl. Learn., vol. 28, no. 2, pp. 611–628, 2023.

[5] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), Florence, Italy, 2019, pp. 6558–6569.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Adv. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008.

[7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.

[8] M. She, M. Xiao, and Y. Zhao, "Technological implication of the digital twin approach on the intelligent education system," Int. J. Humanoid Robotics, vol. 20, no. 3, pp. 2350015, 2023.

[9] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the digital twin: A systematic literature review," CIRP J. Manuf. Sci. Technol., vol. 29, pp. 36–52, 2020.

[10] M. Furini, V. Freschi, A. M. Vegni, and M. C. Vuran, "Digital twins and artificial intelligence as pillars of personalized learning models," Commun. ACM, vol. 68, no. 2, pp. 52–60, 2025.

[11] A. Saluja, N. Baig, A. Grover, and R. Adlakha, "Leveraging digital learning environments and predictive analytics to develop adaptive early warning systems for at-risk students in higher education," in Transforming Operational Efficiency and Strategic Decision Making in Higher Education. Hershey, PA, USA: IGI Global, 2026, pp. 210–228.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnnostic meta-learning for fast adaptation of deep networks," in Proc. Int. Conf. Machine Learning (ICML), Sydney, Australia, 2017, pp. 1126–1135.

[13] A. G. Khoee, Y. Yu, and R. Feldt, "Domain generalization through meta-learning: A survey," Artif. Intell. Rev., vol. 57, no. 1, pp. 1–39, 2024.

[14] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 5149–5169, 2022.

[15] M. Abdollahzadeh, T. Malekzadeh, and N.-M. Cheung, "Revisit multimodal meta-learning through the lens of multi-task learning," arXiv preprint arXiv:2110.14202, 2021.

[16] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learn. Individ. Differ., vol. 103, p. 102274, 2023.

[17] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.

[18] Y.-Z. Lin, A. H. J. Alhamadah, M. W. Redondo, K. H. Patel, et al., "Transforming engineering education using generative AI and digital twin technologies," arXiv preprint arXiv:2411.14433, 2024.

[19] M. Á. Rodríguez-Ortiz, J. M. Dodero, and M. S. Ibarra-Sáiz, "Machine learning and generative AI in learning analytics in higher education: A systematic review," Appl. Sci., vol. 15, no. 15, p. 8679, 2025.

[20] G.-G. Lee, L. Shi, E. Latif, Y. Gao, et al., "Multimodality of AI for education: Towards artificial general intelligence," arXiv preprint arXiv:2312.06037, 2023.

[21] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," arXiv preprint arXiv:1701.06538, 2017.

[22] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning," in Proc. Int. Conf. Machine Learning (ICML), Stockholm, Sweden, 2018, pp. 1568–1577.

[23] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," arXiv preprint arXiv:1803.02999, 2018.

[24] S. Doroudi, E. Brunskill, and M. Brunskill, "Where's the reward? A review of reinforcement learning for educational technologies," Int. J. Artif. Intell. Educ., vol. 29, no. 4, pp. 568–620, 2019.

[25] N. C. Kriegeskorte and P. K. Douglas, "Interpreting encoding and decoding models," Curr. Opin. Neurobiol., vol. 55, pp. 167–179, 2019.

[26] H. ALLIOUI and Y. Mourdi, "Artificial intelligence and education: A systematic literature review," Expert Syst. Appl., vol. 233, p. 120500, 2023.

[27] H. ALLIOUI, Y. Mourdi, and I. Oumaira, "Deep learning and educational data mining for student performance prediction: A comprehensive review," IEEE Access, vol. 12, pp. 45211–45239, 2024.

[28] H. ALLIOUI and Y. Mourdi, "Digital twins in intelligent educational ecosystems: Opportunities, challenges, and future directions," Sensors, vol. 24, no. 6, p. 1987, 2024.

[29] Y. Mourdi, H. ALLIOUI, and H. Kaa, "Generative artificial intelligence for adaptive learning systems: Emerging perspectives in higher education," Multimedia Tools Appl., vol. 84, no. 4, pp. 10231–10258, 2025.

[30] H. ALLIOUI, Y. Mourdi, and H. Kaa, "Multimodal educational intelligence systems using transformer architectures and explainable AI," IEEE Access, vol. 13, pp. 55221–55247, 2025.