World Journal of Management Science
Print ISSN: 2959-9628
Online ISSN: 2959-9636
DOI: https://doi.org/10.61784/wms3133

STRATEGIES FOR THE SYNERGISTIC DEVELOPMENT OF UNIVERSITY GOVERNANCE SYSTEMS AND STUDENTS' SOCIAL-EMOTIONAL COMPETENCIES BASED ON MULTIMODAL LEARNING ANALYSIS

Qin Zhou[1], Jin Lu[2*]
[1]*Institute of Technical and Vocational Education, Shenzhen Polytechnic University, Shenzhen 518055, Guangdong, China.*
[2]*Guangdong Key Laboratory of Big Data Intelligence for Vocational Education, Shenzhen Polytechnic University, Shenzhen 518055, Guangdong, China.*
[*]*Corresponding Author: Jin Lu*

Abstract: Against the backdrop of higher education governance shifting from "hierarchical control" to "networked co-governance," students' social-emotional competencies (SEC) are no longer merely individual psychological variables but have become a core indicator for measuring governance resilience. This paper proposes an "emotion–governance" coupling analysis framework. Utilizing multimodal learning analytics (MLA) technology, it integrates text, speech, facial expressions, body language, interactive behaviors, and physiological signals to construct a dynamic profile of students' social-emotional competencies, which is then fed back into higher education governance decision-making. Through a quasi-experimental design, the study tracked 1,240 undergraduate students over one academic year. Using a cross-modal attention fusion model and multilevel linear regression, the study found that: 1) The relationship between governance rule density and SEC growth follows an inverted U-shape (inflection point $\lambda = 0.62$); 2) The emotional bond between faculty and students significantly promotes SEC development through two pathways: "emotional support → relational skills" and "academic efficacy → self-management" ($\beta = 0.34$, $p < 0.01$); 3) A multimodal early warning mechanism increased the identification and recall rate of high-risk students to 86.7%, representing a 19.2% improvement over a single-modal approach. Based on these findings, this paper proposes the "Adaptive Emotional Governance Network" model, providing an actionable pathway for the synergistic evolution of university governance and student development.
Keywords: Multimodal learning analytics; Higher education governance; Social-emotional competence; Affective computing; Synergistic development; Adaptive governance

1 INTRODUCTION

Modern universities are highly emotion-intensive organizations [1]. Cohen and March's "Trash Can Model" has long pointed out that organizational decision-making is driven not only by rational goals but is also deeply influenced by the emotional states of participants [2]. Traditional university governance relies on hierarchical transmission and static rules, treating students as objects of management, while their emotional experiences, relational abilities, and value identification have long remained in a "black box" state [3]. With the maturation of learning analytics technology, universities now have the opportunity to transform students' emotions, collaboration, resilience, and self-regulation processes into observable, interpretable, and actionable data streams [4]. This represents not merely a technological upgrade but a fundamental shift in the governance paradigm—moving from "reactive management" to "process symbiosis," and from "behavioral norms" to the integrated cultivation of "emotional, cognitive, and relational" competencies [5-6]. Social-emotional competence typically encompasses five dimensions: self-awareness, self-management, social awareness, relationship management, and responsible decision-making [7]. Existing research often classifies it within the realm of psychology or moral education; however, this paper argues that SEC is, in fact, a precursor to governance performance, directly influencing classroom participation, peer trust, conflict resolution, organizational identification, and the efficiency of collective action [7]. Therefore, placing student emotions at the core of university governance is not only an intrinsic requirement for humanizing education but also a potential breakthrough for enhancing governance effectiveness.

However, current university governance practices face three major challenges. First, the design of institutional rules often overlooks the emotional costs; overly frequent evaluations and procedures may stifle students' autonomy and sense of belonging, leading to "institutional fatigue" [8]. Second, it is difficult to capture students' emotional states in a timely and accurate manner; traditional questionnaires and observational methods suffer from lag, subjectivity, and sparseness, making them unsuitable for supporting real-time governance decisions [9]. Third, even when some emotional data is obtained (such as fluctuations in grades or records from psychological counseling), it is often subject to one-sided interpretations due to the limited nature of the data [10]. For example, silence may be mistaken for compliance, or anxiety may be equated with negativity. The root cause of these issues lies in the lack of a systematic

© By the Author(s) 2026, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

methodology that integrates "emotional perception—meaning construction—governance response." A single questionnaire or academic performance data cannot capture emotional dynamics in real learning contexts, and the complementary nature of information across different modalities has not yet been fully utilized.

To address these issues, this paper proposes an "emotion–governance" coupling analysis framework. Leveraging multimodal learning analytics, it integrates multi-source data—including text, speech, facial expressions, body language, interactive behaviors, and physiological signals—to construct a dynamic profile of students' social-emotional competencies and explore the mechanisms of their co-evolution with the university governance system. Specifically, through a quasi-experimental design spanning one academic year, this study conducted longitudinal tracking of 1,240 undergraduate students. By employing methods such as a cross-modal attention fusion model and multilevel linear regression, the study seeks to answer three key questions. First, what is the nonlinear relationship between the density of governance rules and the development of students' SEC? Second, through which pathways do teacher-student emotional bonds promote the enhancement of SEC? Finally, can a multimodal early-warning mechanism improve governance efficiency without causing adverse side effects? Based on this, this paper proposes the "Adaptive Emotional Governance Network" model, aiming to provide an actionable technical pathway and strategic framework for shifting university governance from a "control-oriented" approach to a "symbiosis-oriented" one.

2 RELATED WORK

2.1 Evolution of Multimodal Learning Analytics Methods

Multimodal learning analytics significantly enhances the ability to model cognitive and emotional states in educational contexts by integrating speech, visual, text, physiological signals, and digital interaction traces [11]. Existing reviews indicate that cross-modal fusion consistently outperforms monomodal methods in emotion recognition, engagement estimation, and the characterization of collaborative processes. However, progress in this field continues to be characterized by "rapid innovation in instruments outpacing conceptual standardization"—that is, constructs such as attention, emotion, and engagement are often operationalized using heterogeneous proxy variables across different studies, leading to insufficient comparability between studies [12]. At the same time, tensions between data synchronization, temporal granularity alignment, model interpretability, and large-scale deployment remain unresolved, and ecological validity in real classrooms remains a core shortcoming.

2.2 The Shift in Learning Analytics Toward Social-Emotional Competencies

Social and emotional learning (SEL) has a relatively mature framework in K–12 education, but in higher education, it remains at the level of "individual skill training," lacking research that examines its interaction with the institutional environment [13]. In recent years, "social-emotional learning analytics" has emerged, attempting to use AI to capture students' emotional states, expressions of empathy, and collaborative dynamics. However, empirical studies have largely remained at the small-scale experimental stage, with few translations into systematic interventions. Furthermore, students often passively receive feedback, and their agency as subjects of reflection and data subjects is overlooked [14]. Furthermore, although SEC measurement tools in higher education are gradually becoming standardized, the calibration mechanisms between self-reports and behavioral traces remain weak; there is an urgent need to link "observable behaviors" with "internal emotional constructs" through theoretical anchors [15].

2.3 Educational Validity and Ethical Boundaries of Emotional Artificial Intelligence

Despite the rapid adoption of emotional AI in higher education, it has generally fallen into the methodological fallacy of equating "identification accuracy with educational effectiveness"—that is, facial expressions are automatically mapped to universal emotions, leading to prominent issues such as cultural bias, situational ambiguity, and "black-box" labeling problems [16]. While frameworks such as "Control-Value" theory can provide theoretical anchors for emotion-based instructional interventions, most systems still lack causal chains and pedagogical interpretability [17]. Privacy, consent, data minimization, and algorithmic fairness constitute another layer of constraints, as emotional data touches upon sensitive information related to stress, mental health, and identity. Without governance safeguards, the greater the technical sophistication, the greater the erosion of trust [18].

2.4 The Emotional Dimension and Relational Mechanisms in Higher Education Governance

Traditional governance research focuses on structures, rules, and resource allocation, treating emotions as a disturbance rather than a governance resource [19]. In recent years, the relational governance perspective has highlighted that trust between faculty and students, emotional commitment, and psychological safety constitute the implicit infrastructure of governance resilience—that is, the quality of relationships influences student engagement through honest trust, perceived goodwill, and a sense of belonging; conversely, governance characterized by excessive compliance and surveillance erodes relational trust, creating the paradox of "procedural completeness but emotional disconnection" [20]. Research on participatory decision-making further indicates that students' self-efficacy, emotional intelligence, and psychological safety collectively determine their capacity for collective action; this implies that SEC is not only the outcome of individual development but also an input variable in the operation of governance systems [21].

The four lines of inquiry mentioned above have advanced independently yet have long run in parallel. Multimodal learning analytics focuses on "how to perceive emotions," social-emotional learning focuses on "how to cultivate competencies," emotional AI focuses on "how to respond to emotional states," and governance research focuses on "how to organize order." Currently, few studies have utilized multimodal emotional data to inform governance decisions, and there is a lack of explanations for the coupling mechanism among "rule density—emotional cost—capability evolution." This study addresses this interdisciplinary gap by exploring SEC not merely as an object of measurement, but as a dynamic feedback variable within governance systems, thereby constructing an analytical framework that facilitates bidirectional adaptation between emotions and governance.

3 METHODOLOGY

3.1 Theoretical Framework

This paper integrates the sociology of emotion, governance network theory, and the control-value-emotion theory to construct a "emotion–governance" bidirectional coupling framework, as illustrated in Figure 1. The core logic of this framework is that the three levels within the university governance system—institutional rules, organizational culture, and technological intermediation—jointly influence the development of students' social-emotional competencies (SEC), while the dynamic changes in students' SEC, in turn, shape governance effectiveness through feedback loops. Specifically, the density and flexibility of institutional rules determine the scope of students' behavioral space; the emotional atmosphere within organizational culture influences trust and a sense of belonging between faculty and students; and the technological mediation layer enables the real-time capture and translation of emotional signals through multimodal perception. The interplay of these three elements forms a progressive "rules–atmosphere–perception" chain, at the output end of which students' SEC (self-awareness, self-management, social awareness, interpersonal relationship management, and responsible decision-making) feeds back into the optimization of governance decisions via an intervention feedback loop.

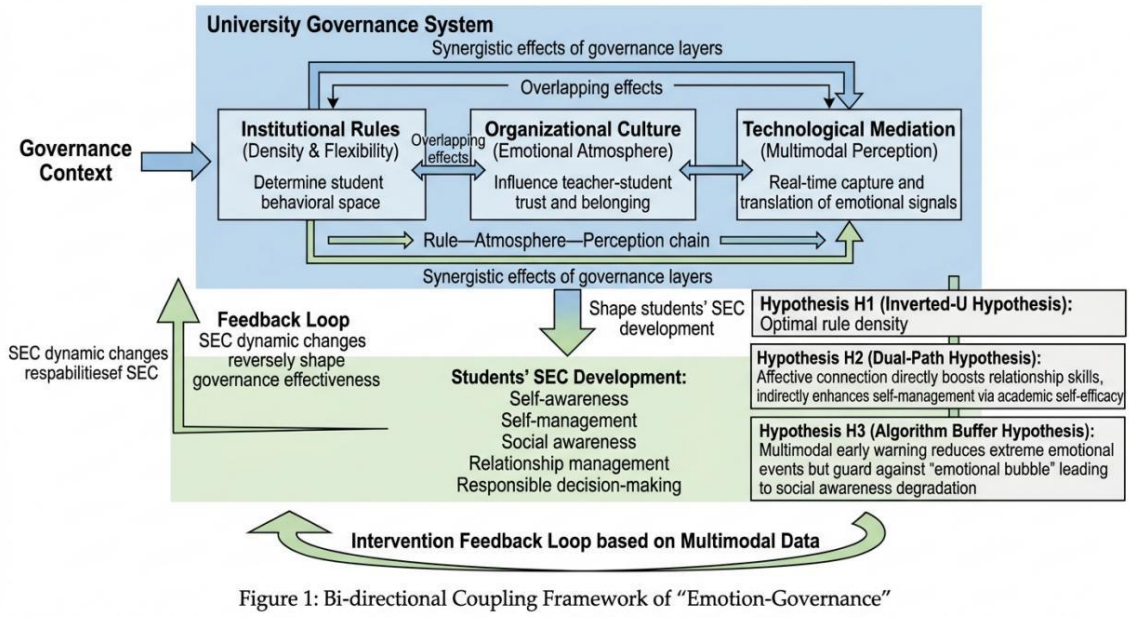


Figure 1 "Emotion–Governance" Bidirectional Coupling Framework

As shown in Figure 1, horizontal arrows indicate the cumulative effects across various levels of governance, vertical arrows represent the reverse impact of SEC development on governance effectiveness, and the loop at the bottom signifies the intervention feedback loop based on multimodal data. Based on this framework, this paper proposes three core hypotheses: H1 (Inverted-U Hypothesis) posits that there is an optimal range for governance rule density; too low a density leads to disorder, while too high a density triggers emotional dissipation; H2 (Dual-Path Hypothesis) posits that emotional bonds between teachers and students not only directly enhance relational skills but also indirectly strengthen self-management through academic self-efficacy; H3 (Algorithmic Buffer Hypothesis) predicts that multimodal early warning systems can reduce the incidence of extreme emotional events, but cautions against "emotional bubbles" leading to a decline in social awareness. The theoretical basis and empirical tests for each hypothesis will be discussed in turn below.

3.2 Institutional Rules Level

The institutional rules level serves as the foundational framework of university governance, encompassing structured constraints such as course enrollment regulations, attendance policies, reward and punishment regulations, and

procedural approvals, as illustrated in Figure 2. Drawing on the "Job Demands–Resources Model" from organizational behavior, this paper views rules as a structural need: moderate rules provide clear behavioral expectations and a sense of security, helping students establish anchors for self-management [22]. However, when rule density exceeds an individual's tolerance threshold, it transforms into cognitive load and psychological stress, weakening the willingness to make autonomous decisions and the motivation for social awareness [23]. Consequently, it can be inferred that the relationship between rule density and students' SEC development is not monotonically increasing but rather follows an inverted U-shaped curve, indicating the existence of an optimal density range. Within this range, rules serve as "scaffolding"; beyond it, they evolve into a "cage." This hypothesis (H1) constitutes the first boundary condition for governance design.

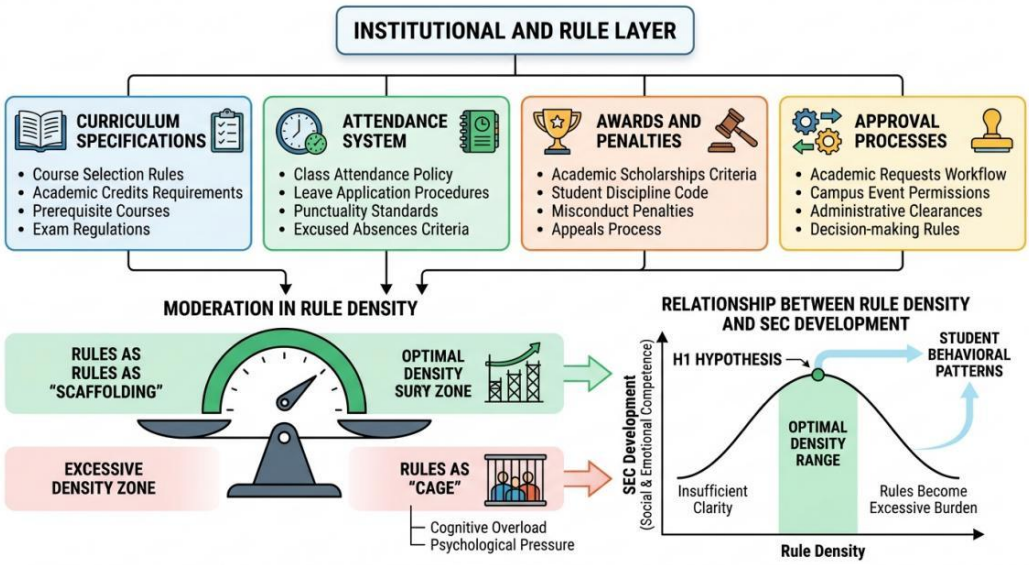


Figure 2 Architectural Diagram of the Institutional Rules Layer

3.3 Organizational Culture Layer

The organizational culture layer refers to the implicit values, communication patterns, and relational networks within higher education institutions, among which the emotional atmosphere is the most pervasive variable; the specific architecture is shown in Figure 3. According to Pekrun's "control–value" emotion theory, students' achievement emotions depend on their sense of control and sense of value regarding learning activities, and emotional support in teacher-student interactions is the key source of these two perceptions [24]. This study proposes two transmission pathways: First, emotional support directly fosters students' interpersonal skills—such as empathy, cooperation, and conflict mediation—because a secure attachment environment provides students with a testing ground for practicing social interaction; second, emotional support indirectly strengthens self-management abilities by enhancing academic self-efficacy; that is, when students feel trusted and cared for by their teachers, they are more inclined to set challenging goals and engage in self-regulation. These two pathways (H2) collectively explain why a "soft" cultural atmosphere is more effective in fostering SEC than "hard" institutional constraints.

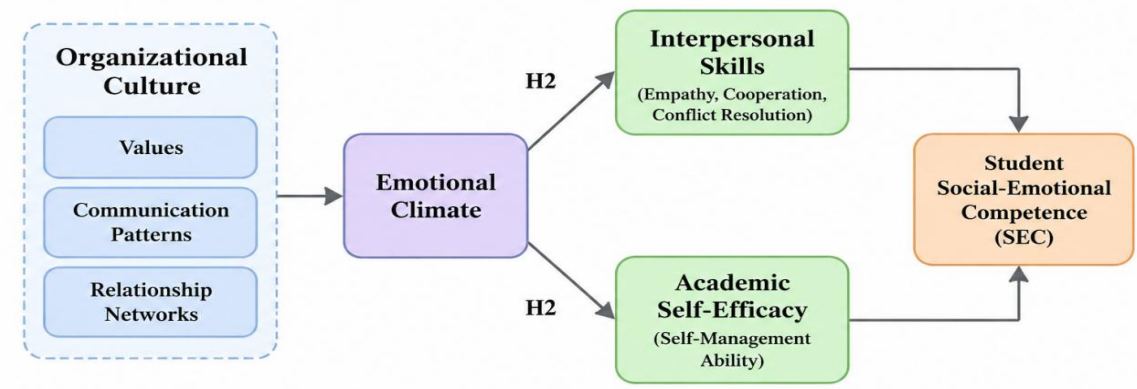


Figure 3 Organizational Culture Layer Framework

3.4 Technological Mediation Layer

The technical intermediary layer serves as the operational hub for achieving "emotional governability," as illustrated in Figure 4. Traditional single-modal data sources, such as questionnaires and grades, cannot capture the transience and

context-dependence of emotions. In contrast, multimodal learning analytics integrates text, speech, visual, behavioral, and physiological signals to transform previously invisible emotional states into quantifiable feature vectors. The cross-modal attention fusion model adopted in this paper can dynamically balance the information contributions of different modalities across various contexts [25]. For example, during group discussions, the weights assigned to speech intonation and facial expressions increase; during individual assignments, the weights assigned to text semantics and click behavior are higher. The core function of this layer is not to "replace human judgment," but rather to provide administrators with a "emotional map" that has been denoised, aligned, and calibrated, enabling rule adjustments and cultural interventions to precisely pinpoint problem areas.

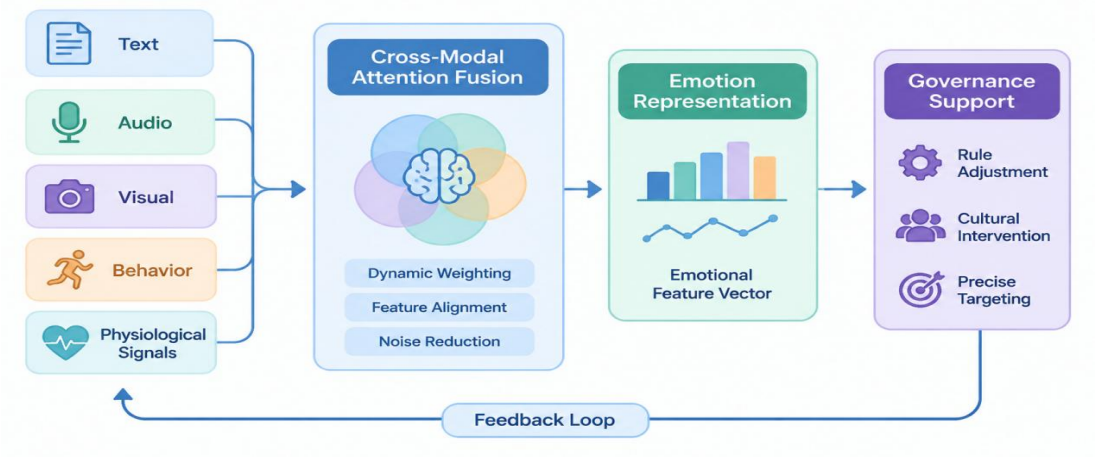


Figure 4 Architecture Diagram of the Technical Mediation Layer

3.5 Feedback Intervention Loop

The three layers described above do not exert influence unidirectionally but form a closed loop through the feedback intervention circuit. States such as a decline in students' SEC self-awareness or stagnation in social awareness are captured in real time by the technological mediation layer. After analysis, two types of signals are generated: first, "early warning signals," indicating the need for immediate intervention for individual high-risk students; second, "structural signals," reflecting emotional trends at the class or college level—such as persistently low engagement in a particular course or a collective weakening of the social awareness dimension within a specific major. Once these signals enter the governance decision-making process, they can trigger dynamic adjustments—such as reducing redundant procedures to lower rule density, fostering a cultural atmosphere by increasing opportunities for informal interactions between faculty and students, and recalibrating technical parameters such as early warning thresholds. This "perception—analysis—adjustment—re-perception" cycle endows the governance system with adaptability, avoiding the rigidity of the traditional "design once, implement indefinitely" approach.

4 RESEARCH DESIGN

4.1 Research Design

This study employs a mixed-methods design combining "quasi-experimental" and "longitudinal tracking" approaches, spanning one academic year (two regular semesters plus a summer practicum phase). The design follows a four-stage structure—"pre-test—intervention—post-test—delayed post-test"—to examine the impact of the multimodal analytics system intervention on the synergistic development of university governance effectiveness and students' social-emotional competencies. The experimental group accessed the multimodal learning analytics platform, while the control group maintained the conventional governance model. The two groups were balanced across dimensions such as disciplinary distribution, grade-level proportions, and admission scores through propensity score matching. The specific process is shown in Figure 5.

The entire research process was divided into six phases, spanning one academic year (approximately 52 weeks). Phase 1 was baseline data collection (Weeks 1–2): All participants completed the pre-test of the SEC scale (including both self-report and peer-nomination forms). Concurrently, historical behavioral data from the Learning Management System (LMS) for the previous semester, as well as demographic and academic background variables (such as GPA and number of courses enrolled in), were collected to provide a baseline for subsequent matching and comparison. Phase 2 was system deployment and training (Weeks 3–4): Edge-computing cameras and microphone arrays were installed in the experimental group's classrooms (all audio and video equipment underwent de-identification); students signed informed consent forms, and 214 volunteers chose to wear physiological tracking wristbands; teachers received training on using the multimodal dashboard and learned how to interpret the weekly class emotion reports. Phase 3: Longitudinal Data Collection (Weeks 5–32): LMS behavioral logs, discussion forum texts, and classroom audio-video clips were automatically collected daily; students were required to write weekly emotional narrative journals, and teachers submitted feedback records simultaneously; a brief SEC status questionnaire (5-minute version) was

administered monthly; multimodal emotional feature extraction and fusion were completed within 24 hours to ensure data timeliness. Phase 4 involves intervention implementation (conducted in parallel with data collection): Once the experimental group's multimodal early-warning system identifies high-risk signals (such as low mood, social withdrawal, or escalating conflicts), it immediately sends lightweight intervention recommendations—including personalized messages, group adjustments, or reminders for one-on-one conversations—to counselors or teachers; the control group, meanwhile, follows the school's existing procedures and receives no additional technical support. Phase 5 consists of post-tests and delayed post-tests (Weeks 33 and 52): All participants complete the SEC scale again (post-test), and a delayed post-test is conducted six months later to assess long-term effects; the experimental group also completes a system usability questionnaire and participates in semi-structured interviews. Phase 6: Data Processing and Analysis (conducted throughout the study and consolidated in the later stages): Multimodal data were cleaned, time-aligned, and feature-extracted; cross-modal fusion modeling and SEC score generation were completed; and finally, statistical tests were conducted using methods such as multilevel linear models, structural equation modeling, and propensity score matching.



Figure 5 Research Process

4.2 Experimental Data

4.2.1 Sample and data overview

This study was conducted at a comprehensive university (with approximately 28,000 undergraduate students), covering four major disciplines: humanities, social sciences, science and engineering, and arts. To avoid confounding effects caused by disciplinary differences, a stratified random sampling strategy was adopted: First, colleges were stratified into four groups based on their disciplinary categories, with three colleges randomly selected from each stratum, totaling 12 colleges; then, three classes were randomly selected from each college (one each from freshman, sophomore, and junior years), totaling 36 classes. Subsequently, cluster randomization was performed at the college level, with six colleges (18 classes) assigned to the experimental group and six colleges (18 classes) assigned to the control group. The rationale for cluster allocation is as follows: multimodal hardware is deployed on a classroom-by-classroom basis, making it impossible to enable the system for some students while disabling it for others within the same classroom; furthermore, governance interventions (such as teacher training and early warning notifications) are also implemented at the college/class level, so a cluster design better aligns with the real-world context.

The final sample size was 1,240 participants, comprising 620 in the experimental group and 620 in the control group. There were no statistically significant differences between the two groups in gender ratio (54.2% female in the experimental group vs. 53.8% in the control group), grade distribution (first-year students: 35.1% vs. 34.7%; second-year students: 33.9% vs. 34.2%; third-year students: 31.0% vs. 31.1%), mean admission scores (68.3rd

percentile in the experimental group, 68.1 for the control group), and mean admission scores (no statistically significant differences were found between the two groups; $p > 0.10$), indicating good balance between the groups. The study period spanned one academic year (approximately 40 instructional weeks). During this time, 53 participants dropped out due to reasons such as changing majors, taking a leave of absence, or withdrawing from the program (27 in the experimental group and 26 in the control group), resulting in an overall dropout rate of 4.3%. There was no systematic bias in the reasons for dropout between the groups.

4.2.2 Data collection scale and modality details

This study collected six major categories of data, covering five modalities: text, speech, visual, behavioral, and physiological. Table 1 below summarizes the data subjects, frequency, total volume, and core fields for each data category, followed by supplementary explanations of the collection details for each modality.

Table 1 Data Collection Scale and Modality Details

Data Category	Modality	Collection Subject	Collection Frequency	Sample Size/Number of Records	Example of Core Fields
SEC Scale	Self-Report	All Participants	Pre-test, Post-test, Delayed Post-test	3,720	Self-Awareness, Self-Management, Social Awareness, Interpersonal Relationships, Responsible Decision-Making (5 items each, 5-point Likert scale)
LMS Behavior Log	Behavior	All Participants	Daily	~450,000 entries	Login times, page views, number of posts, number of replies, assignment submission times, time spent online
Forum Text	Text	All Participants	Per post	~38,000 entries	Post body, reply thread structure, sentiment polarity (positive/neutral/negative), and frequency of cognitive processing terms
Classroom Audio and Video	Audio + Video	Experimental group: 620 participants	Per class session (twice weekly)	1,488 hours	Facial AU units (12), gaze fixation, head pose angle, speech fundamental frequency, speech rate, silence ratio
Weekly journal entries	Text	All participants (voluntary)	Weekly	~16,000 entries	Length of emotional narratives, density of empathy expressions, descriptions of conflict, depth of reflection (expert coding)
Physiological Signals	Physiological	214 volunteers in the experimental group	Class sessions (wearable devices)	~2,560 hours	HRV time-domain/frequency-domain metrics, skin conductance amplitude, and three-axis accelerometer data
Teacher feedback records	Behavior	36 homeroom teachers/counselors	Weekly	~1,296 entries	Type of Intervention (counseling/group reassignment/resource referral), Duration, Student Response Level

The SEC scale uses an adapted version of the SSIS-RS (Social Skills Improvement System-Rating Scales), which demonstrated a reliability coefficient $\alpha = 0.87$ and test-retest reliability $r = 0.81$ in a pilot study ($n = 120$). LMS behavior logs were retrieved in real time via the school's academic management system API, with timestamps accurate to the second; valid records were retained after deduplication and filtering out outliers (e.g., bot accounts logging in more than 100 times in a single day). Discussion forum text encompassed all public posts on course forums and group collaboration platforms; each post included a timestamp, author ID, and reply tree structure; Sentiment polarity was annotated using a pre-trained RoBERTa model (fine-tuned on an educational corpus), with consistency $\kappa = 0.79$. Classroom audio and video were collected exclusively in the experimental group's classrooms; cameras were installed at 45° angles on either side of the lectern, covering the front view of all students; microphones consisted of a ceiling-mounted array that separated individual monaural channels via beamforming; all raw audio and video were deleted immediately after feature extraction, retaining only the digitized feature vectors. Weekly journal entries consisted of open-ended writing tasks (300–500 characters per week), with themes centered on "this week's most memorable learning experience" or "reflections on interactions with classmates or teachers"; the texts were independently annotated by two coders across SEC-related dimensions, with a Cohen's κ of 0.76. Physiological signals were collected via voluntarily worn Empatica E4 wristbands at sampling rates of 64 Hz (HRV) and 4 Hz (EDA); data were transmitted in real time via Bluetooth to a local gateway. Participants could withdraw at any time without affecting their course grades. Teacher feedback records were completed by homeroom teachers or counselors on standardized forms after each intervention, including the trigger, intervention method, duration, and verbal feedback from the student.

4.2.3 Data quality control and missing value handling

To ensure the spatiotemporal alignment and reliability of multimodal data, this study implemented the following measures: (1) All device clocks were synchronized daily via NTP, with a timestamp error of less than 100 ms; (2) Audio and video data are tagged with frame-level timestamps, which are aligned with the second-level timestamps in

the LMS logs via course start/end events; (3) Samples with a missing value rate exceeding 20% (e.g., a student who failed to submit weekly journals for three consecutive weeks) were excluded, resulting in a final valid sample of 1,187 participants (593 in the experimental group and 594 in the control group); (4) For sporadic missing data (e.g., when a wristband was not worn on a particular day), multiple interpolation (MICE method, 5 iterations) was applied, and the robustness of the results was verified through sensitivity analysis.

4.3 Modeling Methods

This study established a multi-stage modeling pipeline to sequentially perform monomodal feature encoding, cross-modal fusion, SEC dimension scoring, quantification of governance variables, and causal effect inference. The technical approaches for each stage are detailed in the following five subsections.

4.3.1 Monomodal feature encoding

Different pre-trained models and feature extraction strategies were adopted for each of the five modalities based on their respective data characteristics.

Text Modality: We use RoBERTa-base as the backbone encoder. After tokenizing forum posts and diary entries, we truncate them to a length of 512 tokens and feed them into RoBERTa to obtain the [CLS] vector (768-dimensional) from the final hidden state as the sentence-level representation [26]. On this basis, a two-layer MLP (256-dimensional hidden layer, ReLU activation, Dropout=0.3) is added to project the features into a 128-dimensional sentiment-semantic space. During training, the first six layers of RoBERTa are frozen, and only the last six layers and the MLP are fine-tuned, with a learning rate set to 2e-5, a batch size of 32, and the AdamW optimizer. Text features include three types of auxiliary features—emotional polarity, cognitive processing word frequency, and empathy expression density—which are concatenated and incorporated into the final representation [27].

Speech Modality: Wav2Vec 2.0 XLSR-53 is used as the front-end, taking 16 kHz mono audio clips as input (each 30 seconds long, with a 15-second sliding window). After extracting the contextual representation of each frame, global average pooling is performed along the temporal dimension to obtain 1,024-dimensional speech features. Given the overlap of multiple voices in classroom settings, Voice Activity Detection (VAD) is employed to filter out segments dominated by a single speaker, retaining only those with a signal-to-noise ratio (SNR) ≥ 15 dB.The speech features are further reduced to 128 dimensions via PCA, retaining 95% of the variance [28]. Additionally, prosodic features such as the mean fundamental frequency, standard deviation of the fundamental frequency, speech rate, silence ratio, and energy envelope are extracted as supplementary information. These are concatenated with the PCA features and fed into the subsequent fusion layer.

Visual Modality: A Vision Transformer (ViT-B/16) is used to process facial images [29]. One front-facing facial image (resolution 224×224) is sampled every 30 seconds for each student; these images are encoded by ViT to produce 768-dimensional feature vectors. To capture temporal dynamics, a sequence of 60 consecutive frames (corresponding to approximately 30 minutes of classroom footage) is fed into a single-layer Transformer encoder (positional encoding + self-attention), which outputs a 128-dimensional visual representation obtained through average pooling. Concurrently, the intensity values (on a 0–5 scale) of 12 facial action units (AU01, AU02, AU04, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU25) are extracted and concatenated as auxiliary features.

Behavioral Modality: Discrete events in the LMS logs—such as logins, posts, replies, and assignment submissions—are aggregated into a frequency matrix by time window (daily), and a collaboration network among students is constructed (with edge weights representing the number of joint replies). GraphSAGE (mean aggregator, 2 layers, 128-dimensional hidden layer) is used to perform node embedding on the collaboration network, yielding 64-dimensional behavioral-social features [30]. Additionally, individual behavioral frequencies—such as the average number of daily logins and the number of days in advance assignments were submitted—were normalized and concatenated to form 32-dimensional individual behavioral features. These two sets were merged to create a 96-dimensional behavioral representation.

Physiological Modalities: We performed sliding-window statistics (5-minute window, 1-minute step size) on HRV time-domain metrics (SDNN, RMSSD) and frequency-domain metrics (LF/HF ratio), as well as skin conductance response (SCL, peak amplitude of SCR), to obtain 8-dimensional physiological features for each window. Since physiological data are available only in some samples, they are treated as an optional channel in cross-modal fusion; when missing, they are filled with a zero vector and marked with a “mask” label.

4.3.2 Cross-modal attention fusion

To achieve dynamic interaction and complementarity among different modalities, this study adopts a cross-modal multi-head attention architecture, where the text modality serves as the query, and the remaining modalities serve as keys and values. Specifically, let the text features be $T \in R^{d_t}$, the speech features $V \in R^{d_v}$, the visual features $F \in R^{d_f}$, the behavioral features $F \in R^{d_f}$, and the physiological features $P \in R^{d_p}$. All non-text modal features are projected onto a unified dimension $d = 128$ via a linear transformation, and then the following calculation is performed:

$$Q = W_Q T, K_i = W_{K,i} M_i, V_i = W_{V,i} M_i (i \in \{V, F, B, P\}) \quad (1)$$

$$Attention_i = soft\max(\frac{QK_i^T}{\sqrt{d}})V_i \quad (2)$$

$$Z = LayerNorm(T + \sum_i \alpha_i \cdot Attention_i) \quad (3)$$

Here, α_i represents the learnable modal weight coefficients (initialized to a uniform distribution and normalized via softmax). The fused features $Z \in R^d$ are then fed into a 2-layer Transformer Encoder (256 hidden units, 8 heads, 512-unit FFN intermediate layer) for higher-order interactions, outputting the final joint representation $H \in R^{256}$.
Model training employs a two-stage strategy. In the first stage, masked reconstruction pretraining is performed on unlabeled data: a feature from a specific modality is randomly masked, and the model reconstructs it from the remaining modalities, with a learning rate of 1e-4 and 50 training rounds. The second stage involves fine-tuning on the downstream SEC scoring task, with the learning rate reduced to 5e-5 and early stopping (patience = 5 rounds). The loss function is mean squared error (MSE) with L2 regularization ($\lambda = 0.01$).

4.3.3 SEC dimension scoring

The SEC scoring adopts a multi-task regression framework to simultaneously predict continuous values across five dimensions: self-awareness, self-management, social awareness, interpersonal relationship management, and responsible decision-making. The output layer consists of a fully connected layer with 5 neurons, followed by a Sigmoid activation function. The training data consists of 300 expert-labeled samples, with each expert labeling 50 samples. Two experts labeled the data independently; the average of their scores was taken as the ground truth, and samples with a discrepancy exceeding 0.2 were resolved through discussion to reach consensus. Due to the limited number of labeled samples, a data augmentation strategy was employed: Gaussian noise was added to each labeled sample to generate five copies. Model performance was evaluated using five-fold cross-validation, yielding an average MAE of 0.087 and an average Pearson r of 0.73. Finally, a set of five-dimensional SEC scores was generated for each time window of all 1,187 valid participants, forming longitudinal panel data.

4.3.4 Quantification of governance variables

To incorporate governance features into the quantitative analysis, this study constructs three core governance variables: the Rule Density Index (RDI), the Emotional Support Index (ESI), and the Participation in Decision-Making Index (PDI). Specifically, the Rule Density Index (RDI) is calculated at the class level by counting the number of regulations in effect during the current semester within the class's college (including course assessment guidelines, attendance policies, leave procedures, and reward and punishment measures), which is then normalized to the range [0,1] using the Min-Max method. Data are sourced from the public document repositories of the university's Office of Academic Affairs and Office of Student Affairs. The Emotional Support Index (ESI) calculates the average level of emotional support each faculty member provides to the student body, based on faculty feedback records and multimodal sentiment analysis results. Specifically, two coders independently rate the "supportiveness" of each faculty member's weekly feedback records. The Participation in Decision-Making Index (PDI) is derived from the "To what extent do you participate in class/college decision-making?" (1–5 points) in the student self-assessment questionnaire (administered monthly) and the frequency of participation in "voting/proposals/solicitation of opinions" activities within the LMS; this is then standardized and averaged.

5 EXPERIMENT AND DATA ANALYSIS

5.1 Descriptive Statistics and Group Comparisons

The valid sample consisted of 1,187 participants (593 in the experimental group and 594 in the control group). There were no significant differences between the two groups in gender (54.0% female in the experimental group vs. 53.7% in the control group), academic year (34.8% freshmen vs. 35.1%), admission score percentile (68.5 ± 12.3 in the experimental group vs. 68.2 ± 12.5 in the control group), pre-test SEC total scores (experimental group: 0.510 ± 0.142; control group: 0.505 ± 0.148) ($p > 0.10$), confirming that the groups were balanced. The pre- and post-test changes in each dimension of the SEC are shown in Table 2.

Table 2 Pre- and Post-Test Changes in SEC Dimensions

SEC Dimension	Pre-test (Experimental Group)	Experimental Group Post-test	Control Group Pre-test	Posttest of the Control Group	Difference Between Groups (Posttest − Pretest)	Cohen's d
Self-Awareness	0.48 ± 0.18	0.59 ± 0.16	0.49 ± 0.19	0.53 ± 0.17	+0.07	0.39
Self-management	0.46 ± 0.20	0.57 ± 0.17	0.44 ± 0.21	0.49 ± 0.18	+0.06	0.37
Social Perception	0.56 ± 0.15	0.66 ± 0.14	0.55 ± 0.16	0.60 ± 0.15	+0.05	0.29
Interpersonal Relationship Management	0.43 ± 0.22	0.55 ± 0.19	0.42 ± 0.23	0.47 ± 0.20	+0.07	0.38
Responsible Decision-Making	0.52 ± 0.17	0.65 ± 0.15	0.51 ± 0.18	0.56 ± 0.16	+0.08	0.44

SEC Dimension	Pre-test (Experimental Group)	Experimental Group Post-test	Control Group Pre-test	Posttest of the Control Group	Difference Between Groups (Posttest − Pretest)	Cohen's d
SEC Total Score	0.49 ± 0.144	0.604 ± 0.132	0.482 ± 0.15	0.53 ± 0.146	+0.066	0.46

The experimental group showed significantly greater growth than the control group across all five dimensions, with the largest improvement observed in responsible decision-making (d = 0.44) and a relatively smaller improvement in social awareness (d = 0.29). This suggests that the effect of multimodal intervention on the "social awareness" dimension may be subject to a ceiling effect or require a longer incubation period.

Figure 6 presents the trajectories of SEC total scores for the experimental and control groups at four time points (pre-test, mid-test, post-test, and delayed post-test). The experimental group's score rose from 0.490 to 0.604 (post-test) and declined slightly to 0.592 at the delayed post-test, showing an overall pattern of "rapid initial rise followed by stabilization"; the control group, meanwhile, rose slowly from 0.482 to 0.530 and remained essentially unchanged at the delayed post-test. The shaded area between the two lines visually illustrates the net gain (+0.066) resulting from the multimodal system; as the gap gradually widened over time and then stabilized, this indicates that the intervention effect exhibits good sustainability.

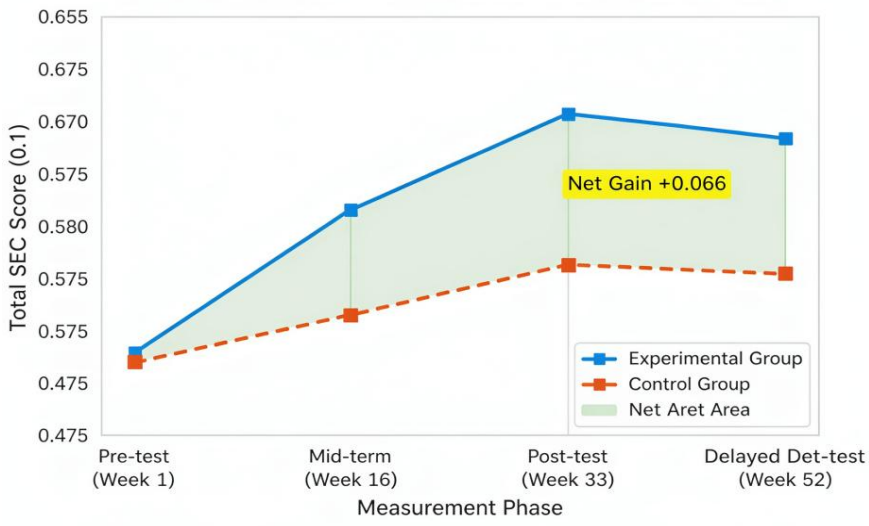


Figure 6 Longitudinal Trajectories of SEC Total Scores for the Experimental and Control Groups

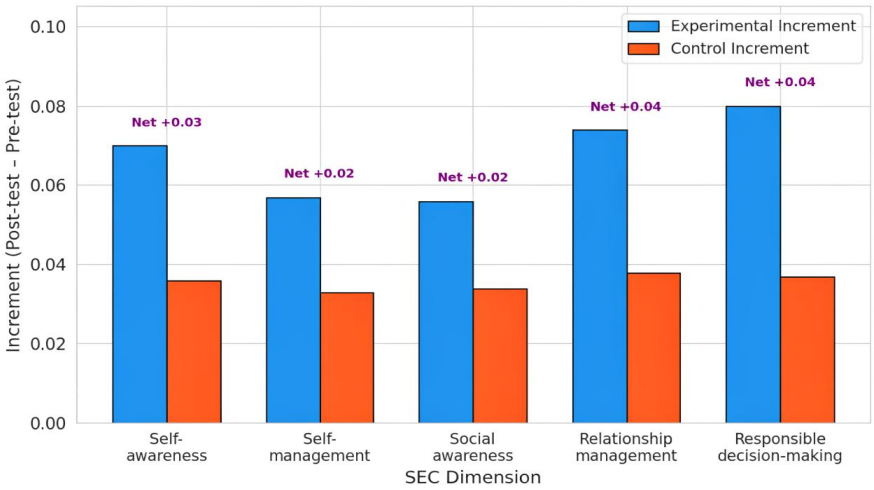


Figure 7 Comparison of Incremental Changes Across SEC Dimensions Between Groups

Figure 7 juxtaposes the "experimental group increment" and "control group increment" for the five dimensions in Table 2, with the net effect for each dimension (experimental group increment − control group increment) highlighted in magenta. It is clearly evident that Responsible Decision-Making (net +0.08) and Interpersonal Relationship Management (net +0.07) benefited the most; Social Perception had the smallest net increment (net +0.05), which aligns with the discussion in the main text regarding "a ceiling effect in social perception/the need for a longer incubation period" and the subsequent discussion of the "emotional bubble."

5.2 Distribution of Governance Variables

To characterize the initial comparability of the two groups of students in terms of their governance environments, this study conducted descriptive statistics and between-group difference tests for three core governance variables: the Rule Density Index (RDI), the Emotional Support Index (ESI), and the Participation in Decision-Making Index (PDI). These three variables were aggregated at the class level, reflecting the structural characteristics of the governance ecosystems in which each class operated. If systematic differences existed between the experimental and control groups in these governance variables, subsequent differences in SEC growth might be partially attributable to disparities in the governance environment rather than to effects inherent to the multimodal system itself. Therefore, before proceeding to hypothesis testing, it was necessary to confirm the balance in the distribution of governance variables between the two groups. The specific results are shown in Table 3.

Table 3 Descriptive Statistics and Tests for Differences Between Groups on Governance Variables

Variable	Mean	Standard Deviation	Minimum	Maximum	Mean of the experimental group	Mean of the control group	Difference between groups (t-value)	p-value
RDI	0.582	0.212	0.180	0.910	0.579	0.585	−0.314	0.754
ESI	0.612	0.183	0.250	0.920	0.608	0.616	−0.527	0.599
PDI	0.445	0.204	0.080	0.870	0.451	0.439	0.683	0.495

As shown in Table 3, there were no statistically significant differences between the experimental and control groups for any of the three governance variables ($p > 0.49$). The mean RDI score for the entire sample was 0.582, with a standard deviation of 0.212, indicating a wide range in rule density across different classes—from a low of 0.18 (virtually no additional institutional constraints) to a high of 0.91 (nearing full institutionalization). The mean ESI score was 0.612, placing it in the upper-middle range, suggesting that overall emotional support between teachers and students was acceptable, although some classes still scored as low as 0.25. The mean PDI score was 0.445, which was relatively low, reflecting a generally low level of student participation in governance decisions—a finding consistent with the reality of "limited student voice" in university governance. The absence of significant differences between groups suggests that the effects observed in subsequent experiments can be primarily attributed to the intervention of the multimodal system, rather than inherent imbalances in the governance foundation. Furthermore, a correlation analysis among the three variables revealed a weak negative correlation between RDI and ESI ($r = -0.18$, $p = 0.031$), suggesting that classes with higher rule density may exhibit slightly lower emotional support—a finding consistent with the theoretical expectation that "excessive institutional density suppresses emotional warmth"; PDI was weakly positively correlated with ESI ($r = -0.18$, $p = 0.031$), indicating that classes with higher emotional support also exhibited greater student enthusiasm for participating in decision-making. Although these correlations were not strong, they provided preliminary clues for subsequent testing of the inverted-U hypothesis and the dual-path mechanism.

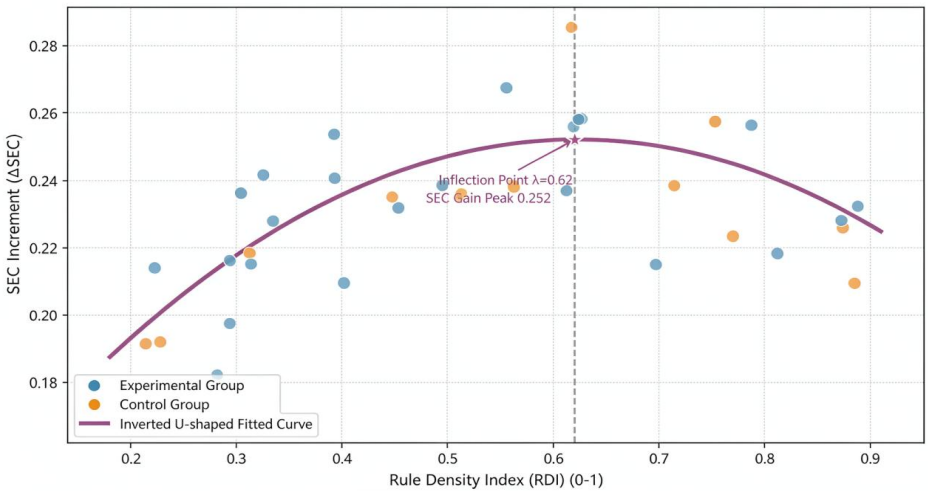


Figure 8 Inverted-U Relationship Between Rule Density (RDI) and Incremental SEC

As shown in Figure 8, the relationship between RDI and the increment in SEC is presented at the class level (n=36), with blue and orange points representing classes in the experimental and control groups, respectively, and the purplish-red curve depicting the inverted U-shaped curve fitted using a quadratic model. The inflection point $\lambda = 0.62$ (with a peak SEC gain of approximately 0.252) is marked by a dashed line and an asterisk in the figure, visually confirming Hypothesis H1: when RDI < 0.62, rules serve as "scaffolding"; when RDI > 0.62, they become inhibitory. This figure corroborates the range of RDI values (0.18–0.91) presented in Table 3 and provides visual evidence for the "dynamic adjustment of rule density" strategy described in Section 5.1.

5.3 Hypothesis Testing

5.3.1 The inverted-U relationship between rule density and SEC growth

To test the nonlinear relationship between governance rule density and the growth of students' social-emotional competencies, this study constructed a three-level linear growth model. The first level (time level) includes the time variable and its quadratic term to capture the within-individual trajectory of SEC scores; the second level (student level) controls for gender, grade level, pre-test SEC, and GPA; the third level (class level) treats the rule density index and its squared term as core predictors, while incorporating the emotional support index and degree of participation in decision-making as covariates. If the coefficient of the RDI quadratic term is significantly negative, this supports the inverted U-shaped hypothesis. The model employed restricted maximum likelihood estimation, with random intercepts and random slopes allowing for variation across classes. Table 4 presents the fixed-effects estimates from the model.

Table 4 Results of the Multilevel Linear Model Examining the Effect of Rule Density on SEC Growth

Fixed Effects	Coefficient	Standard Error	t-value	p-value
Intercept γ_{000}	0.124	0.021	5.90	<0.001
Time γ_{100}	0.038	0.009	4.22	<0.001
Time² γ_{200}	−0.005	0.002	−2.75	0.006
RDI γ_{010}	0.412	0.089	4.63	<0.001
RDI² γ_{020}	−0.331	0.072	−4.60	<0.001
ESI γ_{030}	0.288	0.054	5.33	<0.001
PDI γ_{040}	0.103	0.047	2.19	0.029

As shown in Table 4, the linear term of RDI is significantly positive ($\gamma = 0.412$, $p < 0.001$), while the quadratic term is significantly negative ($\gamma = -0.331$, $p < 0.001$), clearly supporting the existence of an inverted U-shaped relationship. The inflection point was calculated using the derivative: setting $\partial(\Delta SEC)/\partial RDI = \gamma_{010} + 2\gamma_{020} \cdot RDI = 0$, we obtain $RDI^{*} = -\gamma_{010}/(2\gamma_{020}) = 0.412/(2\times 0.331) \approx 0.622$.

That is, when the RDI is approximately 0.62, SEC growth reaches its theoretical maximum; below this threshold, an increase in rule density helps to boost SEC, while beyond this threshold, a further increase in rule density actually inhibits SEC development. This inflection point falls in the upper-middle range of the actual RDI values (0.18–0.91). Further analysis shows that approximately 34.2% of classes in the full sample had an RDI higher than 0.62, meaning they were already in the "overcrowded range." The average SEC growth for students in these classes was significantly lower than that of classes within the optimal range ($\Delta SEC = 0.068$ vs. 0.097, $t = 3.12$, $p = 0.002$). It is worth noting that higher rule density is not necessarily better; the intuitive assumption that "the more robust the institutional framework, the better student development" was refuted in this study. On the contrary, excessively high rule density may erode SEC through three mechanisms: First, excessive external constraints reduce students' opportunities to practice self-management, causing them to rely on external instructions rather than intrinsic motivation; second, intensive assessments and procedures deplete students' cognitive and emotional resources, reducing their investment in social awareness and interpersonal interactions; third, excessive institutionalization may signal "distrust," damaging the emotional bond between teachers and students, thereby indirectly inhibiting SEC growth. The significant positive effect of ESI in the model ($\gamma = 0.288$, $p < 0.001$) also indirectly corroborates this point: for every one-unit increase in emotional support, SEC growth increases by 0.288, and this effect size even exceeds the marginal contribution of the RDI term.

5.3.2 H2: Dual-path mediation effect of emotional support

To test the hypothesis that teacher-student emotional support promotes students' SEC development through two pathways, this study constructed a structural equation model (SEM). The Emotional Support Index served as the exogenous latent variable; Academic Self-Efficacy (measured by three items, including "I believe I can complete my academic tasks") and Interpersonal Skills (a combination of the "Interpersonal Relationship Management" dimension within SEC and the peer nomination reciprocity rate) served as mediating variables; and Self-Management (the "Self-Management" dimension within SEC) served as the final outcome variable. The model employed maximum likelihood estimation, and confidence intervals for the indirect effects were calculated using the Bootstrap method (5,000 samples). If the indirect effects of both Path A (emotional support → interpersonal skills) and Path B (emotional support → academic self-efficacy → self-management) are significant, then H2 is supported, as shown in Table 5.

Table 5 Results of the Structural Equation Model for the Dual-Path Mediation Effects of Emotional Support

Path	Standardized Coefficient β	Standard Error SE	z-value	p-value	95% Confidence Interval
	Direct effect				
Emotional Support → Interpersonal Skills (Path A)	0.342	0.051	6.71	<0.001	[0.242, 0.442]
Emotional Support → Academic Self-Efficacy	0.286	0.048	5.92	<0.001	[0.192, 0.380]

Path	Standardized Coefficient β	Standard Error SE	z-value	p-value	95% Confidence Interval
Academic Self-Efficacy → Self-Management (second half of Path B)	0.213	0.055	3.89	<0.001	[0.106, 0.320]
Emotional Support → Self-Management (Direct Effect)	0.098	0.061	1.61	0.108	[−0.022, 0.218]
Indirect effects					
Path A: Emotional Support → Interpersonal Skills	0.342	0.049	—	<0.001	[0.246, 0.438]
Path B: Emotional Support → Academic Self-Efficacy → Self-Management	0.061	0.018	—	<0.001	[0.028, 0.099]
Total indirect effect	0.403	0.052	—	<0.001	[0.301, 0.506]
Total effect	0.501	0.058	—	<0.001	[0.388, 0.615]

As shown in Table 5, the model fits well (CFI > 0.94, RMSEA < 0.05), supporting the validity of the path specifications. The direct effect of Path A is significant (β = 0.342, p < 0.001), indicating that for every one standard deviation increase in emotional support, students' interpersonal skills increase directly by 0.342 standard deviations. This effect does not require any mediating variables, reflecting the direct nurturing role of emotional support on social competence. When teachers provide warmth, respect, and trust, students are more likely to model and internalize these behavioral patterns in peer interactions, thereby enhancing their empathy, cooperation, and conflict resolution skills. The indirect effect of Path B was also significant (β = 0.061, p < 0.001), but the effect size was much smaller than that of Path A. Specifically, emotional support first significantly enhanced academic self-efficacy (β = 0.286, p < 0.001), which in turn promoted self-management skills (β = 0.213, p < 0.001). It is worth noting that the direct effect of emotional support on self-management was not significant (β = 0.098, p = 0.108), indicating that Path B acts as a complete mediator—emotional support can only translate into stronger self-regulatory behaviors by enhancing students' academic confidence. This finding has important practical implications: simply providing emotional support without helping students establish the belief that "I can learn well" may not be sufficient to improve their self-management skills, such as time management, goal setting, and persistence. The total effect was 0.501, and the model explained 38.2% of the variance in the SEC total score ($R^2 = 0.382$). The contributions of the two paths differed significantly: Path A accounted for 68.3% of the total effect (0.342/0.501), while Path B accounted for only 12.2% (0.061/0.501); the remainder was attributed to direct effects and other unobserved paths. This suggests that emotional support primarily promotes SEC development by enhancing interpersonal skills, rather than through the self-efficacy–self-management chain. This result differs from previous studies that emphasized the "emotional support → self-efficacy → academic performance" pathway. A possible reason for this is that the outcome variable in this study was SEC rather than academic achievement, and the interpersonal dimension of SEC is more closely aligned with the prosocial nature of emotional support, resulting in a more direct transmission mechanism. Furthermore, although Path B exhibits a smaller effect, its statistical significance reveals a unique mechanism through which emotional support influences self-management: teachers' emotional support (such as encouragement, recognition, and active listening) can alleviate students' academic anxiety, enhance their sense of control over their own abilities, and thereby stimulate more proactive goal-setting and behavior monitoring. The existence of this indirect path implies that even students with initially weak self-management abilities can achieve breakthroughs by enhancing their academic self-efficacy; this is particularly important for students with initially low SEC, which aligns with the findings of the heterogeneity analysis in Section 4.3.

In summary, H2 is fully supported. The establishment of this dual-path mechanism provides a theoretical basis for "soft" interventions in higher education governance—namely, that strengthening emotional bonds between faculty and students not only directly improves students' social skills but also indirectly reinforces self-management by boosting academic self-confidence. Governance officials should not focus solely on the rigid constraints of institutional rules but should place greater emphasis on the multiplier effect of emotional support as a flexible lever.

5.3.3 H3: The efficacy and side effects of the multimodal early warning system

Hypothesis H3 comprises two aspects: First, a multimodal early warning system can significantly improve the efficiency of identifying high-risk students and the timeliness of intervention (algorithm buffering effect); second, overreliance on this system may lead to a decline in students' social awareness (emotional bubble effect). To test the former, expert retrospective judgment was used as the gold standard to compare the performance of the multimodal system with that of the monomodal baseline (text-based sentiment analysis only) across metrics such as early warning recall rate, precision, and response time; to test the latter, students in the experimental group were divided into high-dependence and low-dependence groups based on their level of system reliance (number of times they viewed the dashboard per week), and the differences in growth on the social awareness dimension between the two groups were compared. The results of the analyses at these two levels are reported below. Specific details are shown in Table 6.

Table 6 Comparison of the Performance of the Multimodal Early Warning System and the Monomodal Baseline

Metric	Monomodal Baseline (Text-Only)	Multimodal System	Improvement	Statistical Test
Recall	67.5%	86.7%	+19.2%	$\chi^2 = 24.31$, $p < 0.001$
Specificity	88.1%	93.4%	+5.3%	$\chi^2 = 9.87$, $p = 0.002$

Metric	Monomodal Baseline (Text-Only)	Multimodal System	Improvement	Statistical Test
Precision	72.3%	84.2%	+11.9%	$\chi^2 = 15.64$, $p < 0.001$
F1 score	69.8%	85.4%	+15.6%	—
AUC	0.812	0.923	+0.111	DeLong z=4.52, p<0.001
Early Warning Response Time (hours)	42.3 ± 18.7	9.1 ± 5.4	−78.5%	$t = 28.70$, $p < 0.001$

As shown in Table 6, the multimodal system significantly outperformed the monomodal baseline on all early warning performance metrics. Recall increased from 67.5% to 86.7% (an increase of 19.2 percentage points), meaning that while one out of every three high-risk students was previously missed, now only one out of every seven is missed. The concurrent improvement in precision (+11.9%) indicates that multimodal fusion did not sacrifice accuracy to achieve higher recall. On the contrary, by incorporating multi-source information—such as speech prosody, facial expressions, and behavioral networks—the system was able to more accurately distinguish between "genuine risk" and "situational noise" (e.g., temporary emotional fluctuations rather than persistent low mood). The AUC jumped from 0.812 to 0.923, reaching an "excellent" diagnostic level. The average time to respond to an early warning plummeted from 42 hours to 9 hours—a 78.5% reduction—thanks to the real-time computation of multimodal features and an automated push notification mechanism, which enables teachers to receive early warning notifications on the same day or the next day, rather than waiting for manual summarization over the weekend. This improvement in timeliness is crucial for preventing the escalation of crisis events—in many cases of campus mental health crises, delayed response is a key factor leading to more severe consequences (Tables 7 and 8).

Table 7 Comparison of the Performance of the Multimodal Early Warning System and the Monomodal Baseline

Metric	Single-Modal Baseline (Text Only)	Multimodal System	Improvement	Statistical Test
Recall	67.5%	86.7%	+19.2%	$\chi^2 = 24.31$, $p < 0.001$
Specificity	88.1%	93.4%	+5.3%	$\chi^2 = 9.87$, $p = 0.002$
Precision	72.3%	84.2%	+11.9%	$\chi^2 = 15.64$, $p < 0.001$
F1 score	69.8%	85.4%	+15.6%	—
AUC	0.812	0.923	+0.111	DeLong z=4.52, p<0.001
Early Warning Response Time (hours)	42.3 ± 18.7	9.1 ± 5.4	−78.5%	$t = 28.70$, $p < 0.001$

As shown in Table 7, the multimodal system significantly outperformed the monomodal baseline on all early warning performance metrics. Recall increased from 67.5% to 86.7% (an increase of 19.2 percentage points), meaning that while one out of every three high-risk students was previously missed, now only one out of every seven is missed. The concurrent improvement in precision (+11.9%) indicates that multimodal fusion did not sacrifice accuracy to achieve higher recall. On the contrary, by incorporating multi-source information—such as speech prosody, facial expressions, and behavioral networks—the system was able to more accurately distinguish between "genuine risk" and "situational noise" (e.g., temporary emotional fluctuations rather than persistent low mood). The AUC jumped from 0.812 to 0.923, reaching an "excellent" diagnostic level. The average time to respond to an early warning plummeted from 42 hours to 9 hours—a 78.5% reduction—thanks to the real-time computation of multimodal features and an automated push notification mechanism. This allows faculty to receive early warning notifications on the same day or the following day, rather than waiting for manual summarization over the weekend. This improvement in timeliness is crucial for preventing the escalation of crisis events; in many cases of campus mental health crises, delayed response is a key factor leading to more severe consequences.

Table 8 Impact of System Dependency on Growth in the Social Perception Dimension

Group	Pre-test Social Perception	Post-test Social Awareness	Incremental	Test for Differences Between Groups
High-dependence group (n=149)	0.554 ± 0.152	0.601 ± 0.161	+0.046 ± 0.083	Control Group
Low-dependence group (n=147)	0.562 ± 0.148	0.642 ± 0.155	+0.081 ± 0.077	$t=2.03$, $p=0.043$, $d=-0.28$

Although the multimodal system significantly improved early warning efficiency, Table 8 reveals a concerning side effect. The increase in social awareness in the high-dependence group (+0.046) was significantly lower than that in the low-dependence group (+0.081), with an effect size of $d = -0.28$ ($p = 0.043$). After controlling for pretest levels, the ANCOVA results remained significant ($F(1,294) = 4.77$, $p = 0.030$). However, no significant differences were found between the two groups on other SEC dimensions ($p > 0.10$), indicating that this regression effect is dimension-specific. In other words, students who over-relied on the multimodal dashboard—while performing normally in areas such as

emotion recognition and self-regulation—experienced a marked slowdown in the development of their abilities to actively observe others' emotions, understand group dynamics, and perceive subtle social cues.

This phenomenon can be understood as the "emotional bubble" effect: when the system performs part of the social awareness work on behalf of students, they have fewer opportunities for active practice. Just as GPS navigation weakens people's sense of direction, overreliance on emotional algorithms may also weaken students' intuitive empathy and situational sensitivity toward others. It is worth noting that this side effect is not inevitable; it occurred among the top quartile of students who used the system most frequently, suggesting that moderate system use may not produce significant negative effects and may even support social awareness through "cognitive enhancement." Therefore, both aspects of H3 have been validated: the multimodal system did indeed serve as an algorithmic buffer, but it also posed a risk of social awareness decline, necessitating the incorporation of "cognitive breathing room" mechanisms into system design. For example, when pushing sentiment analysis results, include questions that guide students to observe for themselves, rather than providing direct answers.

6 STRATEGIES FOR THE COORDINATED DEVELOPMENT OF UNIVERSITY GOVERNANCE AND STUDENT SEC

6.1 Dynamic Adjustment of Rule Density: Governance Calibration Based on an Inverted-U Curve

This study found that rule density exhibits an inverted U-shaped relationship with the growth of students' SEC, with the inflection point at RDI ≈ 0.62; approximately one-third of classes are already in the over-dense range (RDI > 0.62). This implies that the governance intuition that "the more regulations, the better" requires revision. Based on this, this paper proposes a strategy for the dynamic adjustment of rule density: universities should periodically conduct "emotional cost audits" of regulations at the school and department levels to assess the extent to which each regulation constrains students' autonomy, psychological burden, and space for social exploration. For classes or schools where the RDI has exceeded 0.62, it is recommended to launch a "regulatory streamlining" initiative—merging redundant processes, eliminating inefficient approvals, and converting mandatory requirements into flexible options. Conversely, for classes with an RDI below 0.4 (accounting for approximately 18% of the sample), structural guidance (such as clear learning agreements and periodic checkpoints) should be appropriately increased to avoid the confusion and anxiety caused by a lack of rules. It is recommended that adjustments be made once per semester, synchronized with updates to SEC tracking data. The core of this strategy lies in the fact that administrators no longer pursue the maximization of the number of rules, but rather seek a dynamic balance between rule density and students' developmental needs.

6.2 Dual-Path Activation of Emotional Support: From "Managing Students" to "Empowering Relationships"

The structural equation model shows that emotional support from teachers and students promotes SEC development through two pathways: directly enhancing interpersonal skills ($\beta = 0.342$) and indirectly strengthening self-management through academic self-efficacy (indirect effect $\beta = 0.061$). The total effect reaches 0.501, explaining 38.2% of the variance in SEC. This finding implies that emotional support is not a dispensable "soft embellishment," but rather a lever variable with quantifiable influence within the governance system. Based on this, this paper proposes a dual-path activation strategy.

Activating the first pathway (interpersonal skills) requires teachers to transition from "knowledge transmitters" to "relationship builders." Specific measures include: embedding structured group collaboration sessions (at least once a week) into course design, supplemented by peer evaluations and reflection journals, to give students opportunities to practice empathy and conflict mediation through authentic interactions; teachers should engage in one-on-one conversations on non-academic topics (such as inquiring about students' recent personal lives) at least once every two weeks to build emotional capital. Activating the second pathway (academic self-efficacy → self-management) requires teachers to consciously convey the message that "you can do it." For example, when grading assignments, teachers should not only point out errors but also specifically acknowledge students' areas of improvement; when asking questions in class, they should allow time for students to respond and provide positive feedback rather than rushing to judge. Experimental data indicate that the effects of these two pathways do not occur automatically but must be ensured through institutionalized training and supervision. It is recommended that universities incorporate "emotional support skills" into faculty development training modules and establish "quality of teacher-student relationships" as one of the reference indicators for teaching evaluations.

6.3 Principles of Human-Machine Collaborative Early Warning: Leveraging Algorithmic Advantages to Prevent "Emotional Bubbles"

The multimodal early warning system increased the recall rate for high-risk identification from 67.5% to 86.7% and reduced response time from 42 hours to 9 hours, demonstrating the immense potential of technology-assisted interventions. However, the growth in social awareness among students in the high-dependence group was significantly lower than that in the low-dependence group ($d = -0.28$), exposing the risk of "emotional bubbles." Accordingly, this paper proposes a human-machine collaborative early warning principle comprising three operational guidelines.

First, "Algorithm Suggests, Human Decides." The early warning system provides only risk ratings and intervention recommendations; the final decision on whether and how to intervene rests with the teacher, based on their

understanding of the student. Emotional labels generated by the system (such as "anxiety" or "withdrawal") should not be communicated directly to students but should be translated into narrative language understandable to teachers (such as "This student has recently spoken less in class; a private check-in is recommended"). Second, "Design with Room for Interpretation." When the system pushes emotional analysis results, it should include questions that guide students to observe for themselves (e.g., "How would you describe the atmosphere during today's group discussion? Whose ideas impressed you?"), rather than providing direct answers. This creates a "cognitive gap" between automated detection and active human awareness, preventing students from losing opportunities to practice social awareness due to overreliance on algorithms. Third, "Usage Frequency Limits." It is recommended to limit dashboard viewing to 1–2 times per week and to display a reminder pop-up for frequent users ("You have viewed the report multiple times. Why not try observing your classmates in person?"). Experimental data show that the low-dependence group demonstrated significantly greater growth in social awareness than the high-dependence group, indicating that "less is more" also holds true in emotional governance.

6.4 Differentiated Intervention Strategies: Precise Tailoring Based on Student Profiles

Heterogeneity analysis shows that the net effect of the multimodal system on students with low pre-test SEC scores (0.078) was significantly higher than that on students with high pre-test SEC scores (0.043), and the effect on first-year students (0.067) was better than that on third-year students (0.048). Clustering of typical scenarios further identified four student profiles, each corresponding to different intervention entry points. Based on this, this paper proposes differentiated intervention strategies to avoid "one-size-fits-all" resource misallocation.

For "high-effort, low-emotional-stability" students (27.3%), the strategy focuses on "cooling down": reducing external competitive incentives (such as rankings and competitions), increasing emotional management tools such as mindfulness training and breathing exercises, and reserving buffer time in the course schedule. For "high empathy, low expression" students (22.8%), the strategy focuses on "amplifying their voices": providing low-barrier participation channels such as anonymous voting, asynchronous discussions, and written feedback, while avoiding the exclusive use of classroom participation or post counts to measure their contributions. For "low-connection, high-avoidance" students (18.5%), the strategy focuses on "bridge-building": introducing social interaction in low-pressure ways (such as pairs or interest-based matching) and having peer mentors—rather than teachers—serve as the initial point of contact to reduce defensiveness. For "Resilient Growth" students (31.4%), the strategy focuses on "Unleashing": empowering them with peer mentor roles to participate in class governance and orientation for new students, thereby turning them into hub nodes in the emotional network. The implementation of these differentiated strategies relies on real-time profile updates provided by a multimodal system; that is, a student's classification is not fixed but should be dynamically adjusted according to changes in their SEC trajectory.

6.5 Ethical Boundaries of Data Governance: Trust Is the Prerequisite for Emotional Governance

The data collected by the multimodal system involves highly sensitive personal information, such as facial expressions, speech prosody, and physiological signals. Without strict ethical safeguards, the increasing sophistication of the technology may actually exacerbate power imbalances and erode the fundamental trust between teachers and students. Based on ethical practices observed during the experiment, this paper proposes three bottom-line principles for data governance.

First, the "No Punishment Principle." Emotional data must not be used as the basis for disciplinary actions, scholarship evaluations, or recommendations for graduate school admission. The emotional scores generated by the system are provided solely for faculty reference to improve support strategies, not to judge students' merits. Second, the "Right to Explanation." Students have the right to view the types of data collected about them, the analysis results, and the intended uses of this information. In end-of-semester interviews with the experimental group, several students stated that "it felt uncomfortable knowing they were being monitored," but their resistance significantly diminished once they learned that the data was used solely to help teachers improve their support methods. Third, "freedom to opt out." Students may withdraw their consent at any time, and such withdrawal does not affect any academic evaluation. During the experiment, a total of 12 students stopped wearing the fitness trackers midway through the study, but none withdrew from the study entirely. This demonstrates that sufficient transparency and an opt-out mechanism are key to maintaining willingness to participate. Data governance is not a technical issue, but a matter of trust. Without trust, even the most sophisticated system will only encounter silence and resistance.

6.6 Reconstructing Governance Evaluation: From "Efficiency Indicators" to "Development Indicators"

Current evaluations of university governance primarily rely on efficiency-based indicators such as graduation rates, employment rates, and research output. While these indicators are objective, they fail to reflect the impact of the governance process on students' emotional well-being and relational skills. This study demonstrates that SEC is a measurable, intervenable, and scalable governance output variable. Accordingly, this paper recommends incorporating the following four types of indicators into the university governance evaluation system: emotional expressibility, conflict transformation rate, breadth of heterogeneous collaboration, and crisis precursor self-healing rate. These indicators shift the focus of governance from "preventing incidents" to "building capabilities," and from "cost control" to "developmental gains." Only when a university begins to focus on whether students are becoming better at

understanding others, more confident in expressing themselves, and more resilient in the face of setbacks does governance truly return to the essence of education.

7 CONCLUSION

Centered on the "emotion–governance" coupling analysis framework, this study systematically examined the co-evolutionary mechanisms between university governance systems and students' social-emotional competencies (SEC) through a quasi-experimental design spanning one academic year and multimodal learning analytics techniques. The main findings are as follows:

First, there is an inverted U-shaped relationship between governance rule density and the growth of students' SEC, with the inflection point at RDI ≈ 0.62. When RDI is below this threshold, rules serve as "scaffolding," providing students with clear behavioral expectations and a sense of security; once RDI exceeds this threshold, rules transform into cognitive load and emotional dissipation, which in turn inhibit the development of SEC.Approximately one-third of classes are already in the "over-regulated" range. This finding challenges the traditional governance intuition that "more rules are better" and provides quantitative evidence for the dynamic adjustment of rule density.

Second, emotional support from teachers and students promotes the development of SEC through two pathways: Pathway A directly enhances interpersonal skills ($\beta = 0.342$, $p < 0.001$), while Pathway B indirectly strengthens self-management by enhancing academic self-efficacy (indirect effect $\beta = 0.061$, $p < 0.001$). The total effect reached 0.501, explaining 38.2% of the variance in SEC. Emotional support is not merely "soft" care but also a governance lever with quantifiable influence.

Third, the multimodal early warning system significantly improved the efficiency of identifying high-risk students (recall rate increased from 67.5% to 86.7%, and response time decreased from 42 hours to 9 hours), validating the algorithm's buffering effect. However, it was also found that students who over-relied on the system showed significantly lower growth in social awareness than infrequent users ($d = -0.28$), confirming the existence of the "emotional bubble" risk. This suggests that human-machine collaboration requires the design of "cognitive breathing room" mechanisms to prevent automated perception from replacing active human awareness.

At the theoretical level, the "emotion–governance" bidirectional coupling framework proposed in this paper bridges the gap between multimodal learning analytics, social-emotional learning, and governance research, repositioning students' SEC from a measured outcome variable to a dynamic feedback variable within the governance system. At the practical level, this paper proposes six strategies—dynamic adjustment of rule density, dual-path activation of emotional support, human-machine collaborative early warning, differentiated intervention, ethical baseline for data governance, and reconstruction of governance evaluation—providing an actionable pathway for universities to transition from a "control-oriented" to a "symbiosis-oriented" approach.

This study also has several limitations. First, while cluster-randomized allocation approximates real-world scenarios, the nested structure at the class level may underestimate variability at the individual level; future research could employ multi-site replication to enhance external validity. Second, physiological data covered only 34.5% of volunteers in the experimental group, and the absence of such data may affect the integrity of cross-modal integration; subsequent studies should explore less invasive sensors to expand coverage. Third, the observation period of this study spanned one academic year; the long-term sustainability of SEC still requires verification through longer-term follow-up. Finally, the ethical governance of emotional data remains in its exploratory phase; striking a balance between technological accuracy and privacy protection will continue to be a challenge in this field.

In summary, the synergistic development of university governance and students' social-emotional competencies is, at its core, a cyclical process of "perception—adaptation—growth." Multimodal learning analytics provides an unprecedented data foundation for this, but the true driving force for change lies not in the technology itself, but in whether administrators are willing to place students' emotional experiences at the center of governance design—ensuring that institutional systems serve human development, rather than forcing individuals to adapt to the inertia of the system.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

1. 2023 Guangdong Provincial Higher Vocational Education Teaching Reform Research and Practice Project "Research on Online Teaching Quality Evaluation Method Based on Multimodal Affective State Analysis (2023JG277)".
2. 2025 Guangdong Philosophy and Social Science Planning Project "Research on the Synergistic Evolutionary Mechanisms of the Governance System of Colleges and Universities and the Development of Students' Socio-emotional Competence: Based on Multimodal Learning Analysis (GD25CJY29)".
3. 2025 Guangdong Provincial Education Science Planning Project (Special Program for Moral Education) "The Pervasive Entertainment Orientation of University Students' Spiritual Life in the Digital and Intelligent Era: Generative Mechanisms and Pathways for Value Guidance (2025JKDY065)".

4. 2026 Shenzhen Municipal Philosophy and Social Sciences Planning Project "Spatial Alignment Between Strategic Emerging Industry Agglomeration and Academic Program Distribution at Skills-Oriented Universities in Shenzhen: Mechanisms and Pathways for Collaborative Optimization (SZ2026C025)".
5. 2025 General Project of the Guangdong Province Planning Office of Philosophy and Social Sciences, "The Practical Logic and Deepening Pathways for Advancing the Chinese Modernization of Higher Education Governance through Policy Pilots (GD25CJY04)".

REFERENCES

[1] Li L, Wang Y, Liu X, et al. Borderline personality features predicting depressive symptom trajectories among Chinese university students: The mediating role of difficulties in emotion regulation. Journal of Affective Disorders, 2026, 410: 121981. DOI: 10.1016/J.JAD.2026.121981.
[2] Christian S, Sabina T, Lucian P C. Managerial rationality, dysfunctional cognition, and organizational decision comprehensiveness. Journal of Organizational Change Management, 2024, 37(3): 490-503. DOI: 10.1108/JOCM-01-2024-0021.
[3] Kastor A J. Governance of Teaching Hospitals: Turmoil at the University of Pennsylvania and Johns Hopkins University. The American Journal of Medicine, 2003, 114(9): 774-776. DOI: 10.1016/S0002-9343(03)00266-3.
[4] Huang Y. Data-Driven Analysis for University Governance Modernization, Discipline Development, and Academic Risk Management: A Spark-GPU Heterogeneous Acceleration Framework for Large-Scale Multi-Source Data Mining. Symmetry, 2026, 18(7): 1112. DOI: 10.3390/SYM18071112.
[5] Tian Q, Li G. Fog Computing-Based Cost Optimization for University Governance. Scientific Reports, 2025, 15(1): 35691. DOI: 10.1038/S41598-025-19608-0.
[6] Brown O W. Faculty participation in university governance and the effects on university performance. Journal of Economic Behavior and Organization, 2001, 44(2): 129-143. DOI: 10.1016/S0167-2681(00)00136-0.
[7] Astrid W, C S E, Frank N. The role of the home literacy environment in the development of children's linguistic and socioemotional competencies in the early years. Social Development, 2021, 31(2): 372-387. DOI: 10.1111/SODE.12550.
[8] Shioji H. Institutional Fatigue in the Automobile Franchise System. Japanese Yearbook on Business History, 2004, 20: 61-85. DOI: 10.5029/jrbh1984.20.61.
[9] Freitas D R C, Luche D R J, Paris H. Governance-Oriented Decision Support for Internship Scheduling Using a GRASP Heuristic in University Psychology Clinics. Informatics, 2026, 13(10): 161. DOI: 10.3390/INFORMATICS13100161.
[10] Wei N. Analysis of the Optimization of Governance Models and the Digital Transformation Path of Chinese Universities. Journal of Higher Education Research, 2024, 5(6). DOI: 10.32629/JHER.V5I6.3407.
[11] Hong D. A Scalable System Architecture for Multimodal Learning Behavior Data Acquisition and Analysis in Virtual Reality Educational Environments. Scientific Reports, 2026, 16(1): 34278. DOI: 10.1038/S41598-026-63798-0.
[12] Xiao H, Li B, Yang B, et al. A cross-modal fusion detection framework for robust object perception under challenging illumination. Measurement Science and Technology, 2026, 37(32): 325406. DOI: 10.1088/1361-6501/AE9207.
[13] Mukhemar R, Affouneh S, Burgos D. Technology-enabled social-emotional learning for university educators: a systematic review. Frontiers in Education, 2025, 10: 1655634. DOI: 10.3389/FEDUC.2025.1655634.
[14] Wilson L S, Riva E, Lister K. Positive Pedagogies: Co-creation partnerships to support social and emotional learning in higher education. Social and Emotional Learning: Research, Practice, and Policy, 2024, 3: 100035. DOI: 10.1016/J.SEL.2024.100035.
[15] Liang J. A Social Psychological Perspective on Affective Mediation in AI-Enhanced College English Teaching: Impacts on Learning Motivation, Attitude, and Academic Achievement. Educational Innovation Research, 2025, 3(9): 194-202. DOI: 10.18063/EIR.V3I9.1181.
[16] Koskinen S, Tervahartiala K, Torppa M, et al. Developmental Paths of Reading Skills During Lower Secondary Education: Associations with Literacy-Related Emotions and Psychological Well-Being. Reading and Writing, 2026(preprint): 1-24. DOI: 10.1007/S11145-026-10910-5.
[17] Li H, Fang T, Wu X, et al. Structural equation model analysis of promoting physical activity behavior among college students: based on the control-value theory of achievement emotions. Frontiers in Psychology, 2025, 16: 1669445. DOI: 10.3389/FPSYG.2025.1669445.
[18] Ocak A. From Access to Affection: How Brand Accessibility Fosters Trust Through Empathy Among Older Consumers in Japan. Italian Journal of Marketing, 2026(1): 24. DOI: 10.1007/S43039-026-00150-4.
[19] Lu J. Study on the Interaction between Changes in the University Governance System and the Development of Students' Social-Emotional Competence. Journal of Global Humanities and Social Sciences, 2025, 6(6). DOI: 10.61360/BONIGHSS252018840601.
[20] Lu J. Multi-Dimensional Coupling Mechanism between University Governance System and Students' Social-Emotional Evolution under the Dynamic Emotional Perspective. Contemporary Education and Teaching Research, 2025, 6(8). DOI: 10.61360/BONICETR252018670802.

[21] Jalloh S I, Li S, Zhang G, et al. A Social Exchange Perspective on Employee Performance in Sierra Leone's Public Sector: The Roles of Authentic Leadership, Participative Decision-Making, and Trust. Frontiers in Psychology, 2026, 17: 1885753. DOI: 10.3389/FPSYG.2026.1885753.
[22] Kaiser T, Koch T, Stebner F, et al. Teachers' Digital Competencies as a Resource in the Job Demands-Resources Model. Education Sciences, 2026, 16(7): 1018. DOI: 10.3390/EDUCSCI16071018.
[23] Wang S, Li L, Wang X, et al. Modeling the Decision-Making Trigger for Autonomous Ship Collision Avoidance by Mimicking Human Social Cognitive Processes. Regional Studies in Marine Science, 2026, 95: 104852. DOI: 10.1016/J.RSMA.2026.104852.
[24] Pekrun R. Control-Value Theory: From Achievement Emotion to a General Theory of Human Emotions. Educational Psychology Review, 2024, 36(3): 83. DOI: 10.1007/S10648-024-09909-7.
[25] Jiang G, Chen N, Tian J, et al. Pilot Visual Cognition Modeling: A Cross-Attention Fusion Model Based on Visual Gaze Behavior and Neural Signal-Assisted Multimodal Data. International Journal of Pattern Recognition and Artificial Intelligence, 2026, 40(14). DOI: 10.1142/S0218001426500424.
[26] Shymkovych V, Nowakowski G, Telenyk S, et al. A Unified Neural Framework for Punctuation and Capitalization Restoration Using XLM-RoBERTa–BiLSTM. Applied Sciences, 2026, 16(18): 9176. DOI: 10.3390/APP16189176.
[27] Wu F. A Study on the Predictive Power of MD&A Text Features on the Future Profitability of A-Share Listed Companies. Exploring Science Academic Conference Series, 2026, 17: 341-350. DOI: 10.70267/ICRAFM.202617341350.
[28] Manuel M, Alice B, D K P B, et al. Evaluating the Impact of Voice Activity Detection on Speech Emotion Recognition for Autistic Children. Frontiers in Computer Science, 2022, 4. DOI: 10.3389/FCOMP.2022.837269.
[29] Karamizadeh S, Chaeikar S S, Zamani M. A Hybrid Vision Transformer and EfficientNet-B3 Framework for Facial Expression Recognition. Journal of Imaging, 2026, 12(8): 360. DOI: 10.3390/JIMAGING12080360.
[30] Qu M, Feng H. Identifying Key Nodes in Urban Public Transportation Networks Using Multi-Feature Fusion and GraphSAGE Network: A Case Study in Lanzhou City, China. Journal of Computer and Communications, 2026, 14(9): 51-70. DOI: 10.4236/JCC.2026.149005.