International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 232-241 www.ijaem.net ISSN: 2395-5252

Learning Engagement through Multimodal Data Representation: A Theory-Based Practical Framework

Tian-Tian Zhang , Kai-Liang Zheng

Date of Submission: 05-11-2024 Date of Acceptance: 15-11-2024

ABSTRACT: The current integration and innovation of new artificial intelligence and data science are imparting unprecedented impacts and opportunities for the quality of teaching and the digital transformation of education. Consequently, there is a growing focus on the "learning engagement". Learning engagement, which measures the extent of a learner's cognitive, emotional, and behavioral involvement, is a crucial indicator for assessing academic achievement, learning satisfaction, and the quality of education. However, with the advancement of educational informatization, single data sources and traditional evaluation methods are no longer effective in capturing the multidimensional aspects of learner engagement. Therefore, leveraging multimodal data to integrate visual, auditory, physiological, and behavioral information for comprehensively supporting and assessing learner engagement has become an urgent issue. This paper reviews the evolution of the concept and mechanisms of AI-enabled learning engagement. It elaborates on the research paradigms and technological frontiers from the perspectives of various supporting theories, including embodied cognition theory, educational neuroscience, and self-determination theory. A theoretical framework encompassing four key processes—data collection and feature extraction, algorithm design and training, evaluation and optimization, and application—is constructed. Furthermore, it proposes preventive measures for potential data ethics and security issues, aiming to provide valuable references for the multidimensional development of learners and future educational interventions.

Keywords: multimodal data; learning engagement; framework construction; learning analytics

Learning engagementrefers to the learners' subjective initiative and wholehearted participation in learning activities (Philp & Duchesne, 2016). It not only reflects a learning attitude but also embodies deep involvement and sustained commitment to the learning process. Additionally, it serves as one of the critical indicators for evaluating educational quality, reflecting the learners' activeness and the effective utilization of educational resources. In recent years, learning engagement has been considered an essential topic in educational research by numerous scholars.

By summarizing previous studies, it becomes evident that the conceptualization and theoretical discussions surrounding learning engagement are relatively mature. However, the data representation and quantification remain limited to the development of measurement scales. Although multiple scales have been developed to assess learning engagement, most of them focus on evaluating learners' subjective perceptions and rarely involve analyzing and utilizing multimodal data generated during the engagement process. In terms of scientific and practical applications, this field is still at an elementary stage. This is primarily due to limited methods for collecting and processing learning data, which hinder the full utilization of the rich data resources generated during learning engagement. Furthermore, there is a lack of effective data analysis methods, making it challenging to

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 232-241 www.ijaem.net ISSN: 2395-5252

deeply explore the underlying patterns and trends within the data. Therefore, it is imperative to integrate knowledge from multiple disciplines, strengthen the "sticky fit" between multimodal data and educational theories, and promote in-depth applications of learning engagement as well as the exploration of data-driven paradigms that reveal the essence of learning.

I. INTRODUCTION

The concept of learning engagement was first introduced by educational psychologist Ralph Tyler in the 1930s, signifying the degree of learners' active involvement in learning activities (Reeve, 2012). This idea was further developed after Pace (1980) proposed the concept of Quality of effort in academics. Subsequently, as scholars such as Finn (1989) and Marks (2000) introduced the emotional and behavioral dimensions of learning engagement, it began to gain significant attention among researchers. Scholars have defined and categorized learning engagement from various perspectives, gradually forming a widely recognized three-dimensional theoretical framework. Newmann posited that learning engagement is not merely superficial participation, but rather involves deep engagement by learners on cognitive, emotional, and behavioral levels. Cognitive engagement pertains to learners' deep understanding and reflection on learning content; emotional engagement refers to learners' interest and emotional connection to learning activities, describing their mental effort and active participation; while behavioral engagement is expressed through learners' actions in the learning process.

As research has progressed, scholars have found that learning engagement is closely associated with academic achievement and psychological well-being. For example, Fredricks and McColskey (2012) found that students with high levels of engagement perform better academically and exhibit higher levels of psychological health. Moreover, learning engagement interacts with factors such as learning motivation and strategies, creating a positive cycle of learning. It serves as a mediating variable between various influencing factors and outcomes such as academic achievement and learning satisfaction.

Currently, the representation of learning engagement primarily revolves around scales or evaluations of its three dimensions-behavioral, cognitive, and emotional engagement. For example, behavioral engagement is often assessed through observable actions such as asking questions, participating in discussions, taking notes, and frequency of tool usage. Cognitive engagement is mainly summarized through tools like questionnaires and reflective journals, with assessments largely influenced by the subjective judgments of evaluators and the reliability and validity of the scales. Emotional engagement is more complex to measure, as it involves learners' internal psychological states and emotional experiences. It is typically evaluated indirectly through self-report questionnaires in which learners describe their emotions during the learning process, such as interest, satisfaction, anxiety, and emotional avoidance. With advancements in digital technology, some non-verbal cues can capture learners' emotional expressions, such as eye-tracking, emotion analysis software, and voice data. It is worth noting that measurements of the three dimensions should not be divorced from the holistic concept of learning engagement. Given the limitations of evaluation and data analysis methods, integrating "holistic" approaches across different disciplines and specific educational practices remains challenging. Addressing this issue through the construction of frameworks that combine the specificity of educational contexts and the holistic relationship among different dimensions, while incorporating classic educational theories, could provide effective solutions.

II. THEORETICALFOUNDATION

Modeling learning engagement supported by multimodal data requires the integration of multidisciplinary, cross-domain, and multi-perspective theoretical foundations to address deficiencies in educational evaluation and educational data analysis. This entails drawing insights from disciplines such as psychology, education, computer science, and data science to construct a theoretical framework that captures and integrates multiple learner perceptions, thereby illustrating learners' patterns of information processing.

2.1 Embodied Cognition Theory

The Embodied Cognition Theory, emerging since the 1980s, represents a convergence of fields such as epistemology, cognitive science, neuroscience, computer science, and phenomenology in exploring human cognition and knowledge. This cognition theory, rooted in philosophical foundations, centers on a body-mind unity, emphasizing the pivotal role of the body in cognition. The core idea asserts that cognitive processes are embodied, meaning that individuals themselves play a crucial

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 232-241 www.ijaem.net ISSN: 2395-5252

role in cognition. If the human cognitive process were likened to a computer's operation, the body would represent the hardware, while the mind would be the software; both are mutually dependent in the cognitive process. Human mind and cognitive origins are rooted in the body, constituting an integral part of bodily activity. The mind is inherently connected to the body, a concept termed embodied mind, signifying that the mind does not exist independently of the body. Similarly, the initial cognitive processes are intrinsically intertwined with bodily structure and modes of activity. Bodily movements, perception, and interactions with the environment collectively form the foundation of the cognitive world.

In the digital age of interactive and "intelligent" technologies, such as artificial intelligence, the Internet of Things (IoT), and virtual reality, learners' environments are shifting from serial, single-channel operations primarily based on "click" and "button" interactions to learning modes characterized by human-computer interaction, human-machine dialogue, and intelligent spaces.Numerous studies indicate that the Embodied Cognition Theory offers a highly valuable theoretical foundation and insightful inspiration for educational research and teaching practice (see Figure 1).

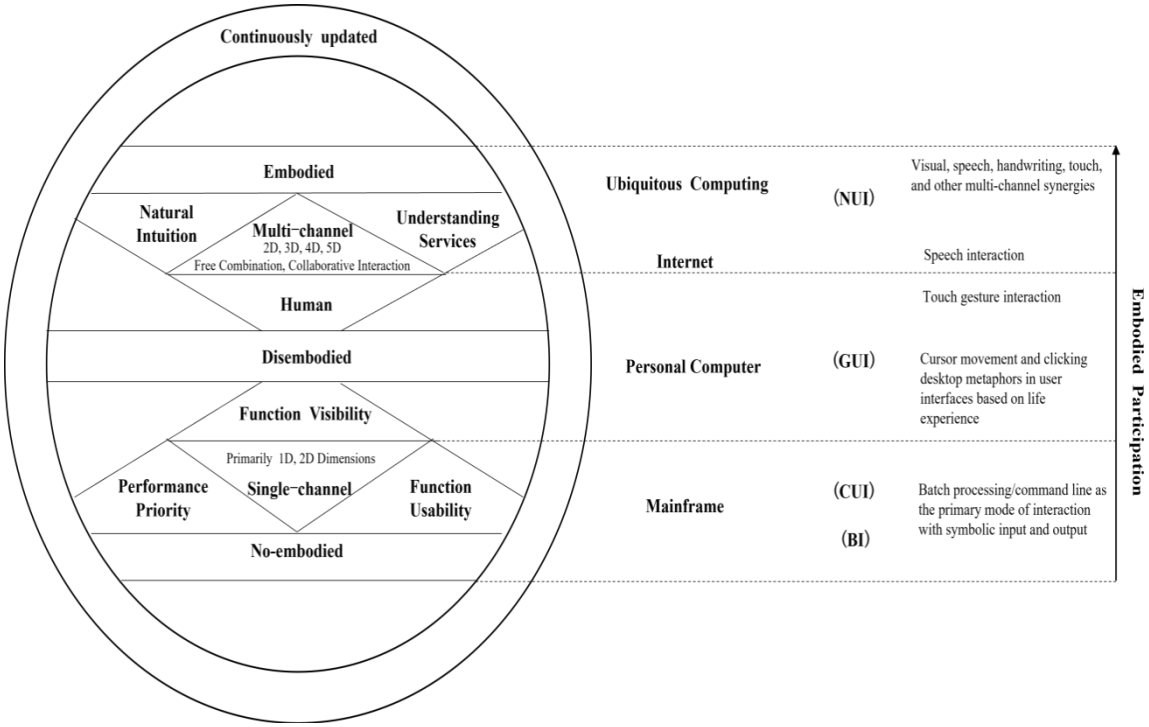

Figure1Embodied Hierarchical Model of Human-Computer Interaction

2.2 Educational Neuroscience

Educational Neuroscience, also known as Neuroeducation, is an emerging research direction that marks a shift from experiential to scientific development within educational research. It represents a natural science paradigm formed by integrating classical education, educational psychology, cognitive science, neuroscience, and information science. Its goal is to provide evidence-based explanations for the underlying mechanisms and neural targets of educational phenomena, serving as a scientific transmitter connecting the neural-psychological-behavioral spectrum, thereby offering new perspectives and methodologies for the study of learning engagement.

Educational Neuroscience emphasizes the corroboration of multi-source and multi-level data, necessitating the integration of multidimensional, dynamic data that can reveal the implicit cognitive states of learners to accurately assess learning behaviors. Previous educational research primarily focused on observable data and multimodal data supported by technology, such as learner dialogue audio, written texts, and videos of learning behaviors, while often neglecting deeper analysis of educational issues and phenomena. Moreover, educational data's "large sample size" and fine-grained "data growth" characteristics often lead to a decrease in data value density, making it challenging to support a deep deconstruction of teaching and learning processes.

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 232-241 www.ijaem.net ISSN: 2395-5252

However, with the advent of emerging technologies, lightweight neurophysiological measurement tools now allow for capturing learners' physiological data—such as eye movements, heart rates, and brain signals—for research and analysis. For example, Keskin et al. (2023) emphasize using eye-tracking technology to visualize teachers' gaze behavior, serving as a reflective tool for educators to better understand their teaching practices. In an experiment by Nafjan and Aldayel (2022), EEG signals were employed to detect learners' attention levels during online courses. By delving into the neural mechanisms underlying learning processes, this research offers a shift from experiential to scientific approaches for enhancing students' focus, engagement, and learning outcomes.

Through understanding the regulation of attention, mechanisms of working memory, and active brain states, Educational Neuroscience provides theoretical guidance for educational practice. Strategies for strengthening learning through pathways such as motivation, emotional regulation, and social interaction stem from a deep comprehension of brain activity, offering critical insights for accurately characterizing learning engagement and improving learning outcomes.

2.3 Self-determination Theory

Self-Determination Theory (SDT), proposed by Deci and Ryan, is a psychological framework aimed at describing the intrinsic nature of human motivation and behavior. SDT's meta-theory posits that all humans are inherently inclined to grow, confront challenges, and voluntarily integrate new experiences. However, these internal tendencies do not operate in isolation; they require an external catalyst. Behavior emerges when internal and external needs are satisfied, driven by motivation rooted in needs. SDT emphasizes three innate psychological needs: autonomy, competence, and relatedness.

The need for autonomy focuses on individuals' freedom to choose and control their own actions, allowing them to act according to their own will, thus becoming masters of their behavior. The need for competence involves experiencing effectiveness in interactions with the environment; when individuals feel capable and can complete tasks, they experience a sense of competence, enhancing intrinsic motivation. This emphasizes individuals' perception of their abilities during these processes. The need for relatedness, or connection, refers to individuals' desire to form meaningful bonds with others and feel like part of a group. Fulfillment of the need for relatedness enhances intrinsic motivation as individuals feel accepted and supported within their social environments. When the external environment offers warm, supportive interpersonal relationships and promotes social connections, individuals' need for relatedness is met.

With the interdisciplinary fusion and in-depth development of educational practices, SDT has increasingly been applied to evaluate learners' mental, behavioral tendencies, and academic achievements within various contexts. From an SDT perspective, learners' behaviors in a dynamic learning process should not merely be passive compliance but instead involve proactive monitoring and adjustment of behaviors based on real-time data to achieve targeted learning goals. When learners' needs for autonomy, competence, and relatedness are met within educational settings, they not only perform better academically but also exhibit higher self-efficacy, more positive emotional states, and sustained interest in learning. These positive emotional and cognitive experiences further enhance learning engagement, creating a virtuous cycle.

III. ALGORITHMICIMPLEMENTATION

Multimodal machine learning aims to learn from heterogeneous yet interconnected data for modeling and prediction (Baltrušaitis et al., 2018). Unlike unimodal models, multimodal models have the unique capability of simultaneously processing and understanding data from diverse sources such as student text submissions, images, and audio communications, thereby enabling more complex and accurate analysis and decision-making. Moreover, multimodal models can use information from one modality to complement or correct incomplete or ambiguous information in another modality, enhancing overall robustness and accuracy. This cross-modal coordination capability makes multimodal models broadly applicable and highly effective within educational applications.

The purpose of multimodal representation is to map data from different modalities into a unified space, facilitating more efficient data processing. The research on multimodal learning did not begin only in recent years; its origins can be traced back to the 1970s. With advancements in deep learning, especially the emergence of large-scale pre-trained models based on Transformers (Vaswani et al., 2017), the effectiveness of multimodal learning has significantly improved, driving its rapid development.

The overall approach is analogous to the

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 232-241 www.ijaem.net ISSN: 2395-5252

training process of many deep learning models. Initially, multimodal data (e.g., text and images) are individually encoded, extracting features relevant to each modality. These feature vectors are then fused, resulting in combined representations, such as text vectors, image vectors, and fused vectors. During the feature fusion process, external knowledge such as graph knowledge, situational knowledge, and domain-specific knowledge can also be incorporated to help the model better understand and process data, thereby improving its generalization performance and robustness. Finally, a loss function is employed to measure the difference between the model's predictions and the true labels, serving as the objective for model optimization.

In the feature extraction stage, text data are often processed using architectures based on Transformers or recurrent neural networks (RNN), while image data are typically encoded using convolutional neural networks (CNN) or methods like Vision Transformers (ViT) that divide images into patches for direct encoding. The key challenge in the feature fusion stage is how to effectively integrate different modalities' features to produce richer and more expressive representations. Currently, single-stream and dual-stream frameworks are the main approaches for multimodal learning fusion. A single-stream model processes all input modalities through one unified flow, typically using shared parameters to facilitate cross-modal information integration and learning. This approach is suitable for data where different modalities are highly correlated, such as text-image matching tasks. Conversely, dual-stream models utilize two parallel flows that separately process different modalities' data before merging their outputs in a particular manner, often employing parallel or sequential structures. Dual-stream models are more effective for heterogeneous data such as text and audio in emotion recognition tasks.

For different data modalities, various loss function techniques are used. For example, text data may use masked language modeling (MLM), where the model predicts masked parts of input text, learning contextual information for words or subwords. For image data, masked region modeling (MRM) masks certain image regions, with the model tasked to predict these concealed areas. When handling text-image data, methods like image-text matching (ITM), image-text contrastive learning (ITC), and image-text global representation (ITG) are applied to align and compare multimodal data.

Table1Comparison of Multimodal Deep Learning Models

Task	Model	Infrastructure	Modal	Dataset	Highlights
Image description	GET (Ji et al., 2021)	Faster-RCNN、LSTM	text、image	MS-COCO	1.In the decoder, a global gated adaptive controller is utilized to integrate relevant information. 2.Inter-layer and intra-layer representations are adopted when merging local and global information.
	VSR (Chen et al., 2021)	Faster-RCNN、LSTM、SSP	text、image	MS-COCO、Flickr30K	1.The semantic roles of specific verbs control the process of image captioning. 2.A human-like semantic structure uses SSP to sort verbs and semantic roles.
	MGAN (Jiang et al., 2021)	Faster-RCNN、LSTM	text、image	MS-COCO	1.Self-gating and attention weight gating are combined with existing self-attention mechanisms to extract relationships within objects. 2.A-pre-layer

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

	CLIP (Radford et al., 2021)	ResNet、Transformer	text、image	WebImageText	normalization Transformer is designed to enhance features. 1.Contrastive learning is used for training. 2.Zero-shot classification tasks can be performed.
Video description	SemSynAN (Perez-Martin et al., 2021)	2D-CNN、3D-CNN、LSTM	video、text	MSVD、MSR-VTT	Guiding language models by fusing syntactic, visual, and semantic representations to enhance the accuracy of descriptions.
	V2C (Fang et al., 2020)	ResNet-LSTM、LSTM	video、text	V2C、MSR-VTT	1. Generating commonsense subtitles for events detected within videos. 2. The V2C-transfer mechanism is capable of producing subtitles enriched with common sense knowledge.
	DPRN (Xu et al., 2020)	MDP、LSTM	video、text	MSVD、MSR-VTT	1. The optimization of generated descriptions is achieved through word denoising and grammatical checking networks. 2. Descriptions are enhanced by taking into account global representations.
Speech synthesis	Parallel TACOTRON (Elias et al., 2020)	GLU、VAE、LSTM	text、audio	Proprietary speech	The model exhibits a high degree of parallelizability during both inference and training processes.
	PC-AVS (Zhou et al., 2021)	GAN、ResNeXt50	text、audio、image	VoxCeleb2、LRW	High-quality synchronized lip-reading of speaking faces is generated through the given facial identity recognizer.
Other tasks	DCWS-RNNs (Lai et al., 2020)	3D-CNN、GRU	text、audio、video	IEMOCAP、AVEC	1. Utilizing four RNNs with context windows of varying sizes to transform context into memory. 2. Multiple graphs based on attention mechanisms integrate these memories for emotion recognition.
	DSCMR (Zhen et al., 2020)	VGGNet	image、text	Wikipedia、XMediaNet、Pascal	1. Minimizing the heterogeneity gap between various

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 232-241 www.ijaem.net ISSN: 2395-5252

			Sentence 、NUS WIDE-10k	modalities. 2. Networks with weight-sharing constraints capture correlations between text and images.
ImageBind (Girdhar et al., 2023)	Transformer	image、text、audio、depth map, heat map, motion vector map	AS-A、ESC、Clotho 、AudioCaps、VGGS 、SUN-D 、NYU-D 、LLVIP 、Ego4D	A joint embedding space was trained.

Table 1 compares some recently proposed multimodal learning models for various tasks, detailing their architectures, data modalities handled, datasets used, and standout features. One of the prominent examples is the CLIP model, which encodes text and image data into vectors using a text encoder and an image encoder, respectively. All data used by CLIP originate from the web, without the need for manual labeling. The model is trained using contrastive learning, a process that brings the vectors of matching text-image pairs closer together in vector space, while pushing apart those of mismatched pairs. A key advantage of CLIP is its ability to perform zero-shot classification tasks. This means that even if the model has not seen an image of an elephant during training, it can still correctly identify it during inference. This capability exemplifies a significant benefit of multimodal models over traditional ones. CLIP represents a breakthrough in visual and language comprehension, offering a novel way to understand the connection between images and language through joint training. Its impact extends beyond impressive performance on various visual and language tasks; it demonstrates a new methodology in learning generalized and robust visual-linguistic representations via contrastive and zero-shot learning approaches from large-scale data. This methodology opens new opportunities for creating more intelligent and adaptive educational AI systems.

As multimodal learning continues to evolve, an increasing number of models aim to integrate more types of modalities. ImageBind exemplifies this by combining six modalities: image, text, audio, depth maps, thermal images, and motion vectors. ImageBind employs pairwise training of these modalities, forming five dual-modal models, effectively aligning the six data types into a shared representation space. This structure grants ImageBind exceptional cross-modal retrieval capabilities-for instance, enabling searches across images, audio, depth maps, and motion vectors using text input. Additionally, ImageBind excels at cross-modal generation, such as generating an image of a pigeon when provided with a pigeon's sound or an initial image. The model's cross-modal computational abilities allow unique interactions; for example, by combining an audio clip of a barking dog and an image of a person walking, the model can generate representations associated with walking a dog. ImageBind's strengths across modalities highlight the immense potential of joint multimodal training for complex and diverse AI applications.

One key advantage of multimodal joint training is its ability to achieve transfer learning across modalities, enhancing the performance of weaker modalities. Strong modalities, such as images or text, typically carry rich information, whereas weaker modalities, such as sound or depth maps, have comparatively less information. Through joint training with strong modalities, weaker modalities can learn enriched feature representations, thereby improving their overall performance. This capacity for cross-modal transfer learning significantly increases the utility and adaptability of multimodal models, making them more effective and robust across diverse applications and scenarios.

IV. FRAMEWORK CONSTRUCTION

4.1 Components of Learner Engagement

The mechanism behind learning engagement is a complex, multi-level process. To construct a theoretical framework for learning engagement, it is crucial to first clarify its underlying mechanisms. Building on theories such as embodied cognition, neuroscience, and self-determination

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 232-241 www.ijaem.net ISSN: 2395-5252

theory, the components of learning activities should be analyzed. These include elements such as learners, educators, learning peers, learning tasks, teaching environments, tool usage, and technological support. Next, it is important to address the measurement dimensions and data representation methods for learning engagement. This challenges the traditional "singularity" of measurement and integrates the use of technology by adopting multimodal data representations, such as images, sound, animations, and videos. It is essential to consider both the feasibility and scientific rigor of the measurement approaches to ensure a comprehensive and accurate understanding of learning engagement.

4.2 Framework Construction for Learning Engagement

By integrating multiple sources of learner data, learning analytics moves beyond relying on a single type of measurement approach. Multimodal data combine traditional subjective measurement methods with physiological measurements (such as facial expressions, voice tone, and body posture), enhancing the accuracy of data analysis and dynamically monitoring the learning process. This leads to a more objective and comprehensive evaluation system. In previous research, many scholars have developed analytical frameworks based on learning engagement to assess learner involvement in specific academic domains. For example, Newell (1992) and Anderson (2002)introduced the concept of Bands of Cognition, exploring multiple dimensions of learning behavior. They divided learning behaviors into biological, cognitive, rational, and social aspects, providing a multidimensional framework to better understand learning behavior. This framework offers a systematic approach to analyzing the various factors that contribute to learning engagement. In terms of group-level engagement features, aggregating and comparing the learning behavior data of different groups can reveal the overall level of engagement and identify common issues or characteristics within the group. This analysis allows for a more comprehensive understanding of learner engagement at both individual and group levels. It can be observed that frameworks for analyzing learning engagement with multimodal data representation typically involve four main processes: data collection and feature extraction, algorithm design and training, evaluation and optimization, and application. By following this general process for multimodal data analysis, this study constructs a learning engagement framework that incorporates multimodal data representation, as shown in Figure 2. This approach provides a more holistic and dynamic understanding of learner engagement, allowing for better support and intervention strategies in the learning process.

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 232-241 www.ijaem.net ISSN: 2395-5252

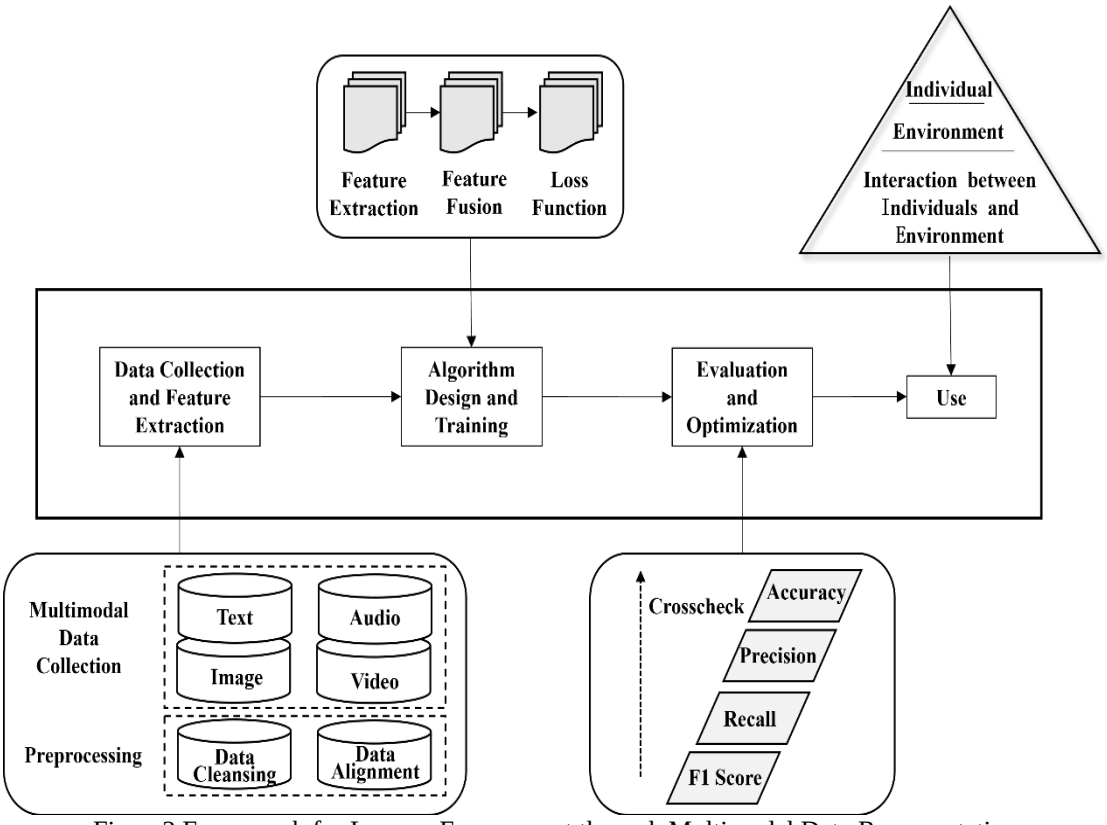

Figure2 Framework for Learner Engagement through Multimodal Data Representation

V. CONCLUSION

Through retrospective research, it is found that evaluating learners' engagement with multimodal data representation is reshaping the learning analytics paradigm, promoting the integration of digital technologies into classrooms and the development of comprehensive evaluation systems. These data can be processed and analyzed using advanced machine learning algorithms and deep learning techniques, revealing fluctuations in learners' cognitive, emotional, and behavioral expressions, measuring their correlation with learning performance, and improving the precision and depth of learning analytics.However, ethical concerns, such as data privacy and security, must be addressed. While ensuring the effectiveness of multimodal data-based assessments, attention must also be given to the anonymity during data collection, access control during data processing, and legal compliance in data storage and analysis. Additionally, future practical applications of the multimodal data representation framework in educational activities should enhance sensitivity to data and foster awareness of risk assessment.

This study systematically reviews the foundational theories of multimodal data and the concept of "learning engagement," and by comparing and analyzing current mainstream multimodal learning algorithms, it explores their advantages in different data representations. Based on this, a multimodal data representation framework for learning engagement is proposed, extending the flexibility of various data modalities in real-world teaching scenarios across four components: data collection and feature extraction, algorithm design and training, evaluation and optimization, and application. This research offers a framework guide from the theoretical and algorithmic calculation perspective. Future research could focus on measuring the dynamic relationships between cognitive, emotional, and behavioral data, including physiological data like eye movement, blood pressure, and iris patterns, within the "learning engagement" concept. This would significantly contribute to promoting learners' abilities and improving the effectiveness of teaching.

REFERENCE

[1]. Al-Nafjan A, Aldayel M. Predict students' attention in online learning using eeg data[J].

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).

Sustainability, 2022, 14(11): 6553.
[2]. Anderson J R. Spanning seven orders of magnitude: A challenge for cognitive modeling[J]. Cognitive Science, 2002, 26(1): 85-112.
[3]. Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2): 423-443.
[4]. Chen L, Jiang Z, Xiao J, et al.Human-like Controllable Image Captioning with Verb-specific Semantic Roles[J]. 2021.
[5]. Elias I, Zen H, Shen J, et al.ParallelTacotron: Non-Autoregressive and Controllable TTS[J]. 2020.
[6]. Fang Z, Gokhale T, Banerjee P,et al.Video2Commonsense: Generating Commonsense Descriptions to Enrich Video Captioning[J]. 2020.
[7]. Finn J D. Withdrawing from school[J]. Review of educational research, 1989, 59(2): 117-142.
[8]. Fredricks J A, McColskey W. The measurement of student engagement: A comparative analysis of various methods and student self-report instruments[M]//Handbook of research on student engagement. Boston, MA: Springer US, 2012: 763-782.
[9]. Girdhar R, El-Nouby A, Liu Z, et al. Imagebind: One embedding space to bind them all[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 15180-15190.
[10]. Ji J, Luo Y, Sun X, et al.Improving Image Captioning by Leveraging Intra- and Inter-layer Global Representation in Transformer Network[C]//National Conference on Artificial Intelligence.2021.
[11]. Jiang W, Li X, Hu H, et al.Multi-Gate Attention Network for Image Captioning[J].IEEE Access, 2021, PP(99):1-1.
[12]. Keskin Ö, Seidel T, Stürmer K, et al. Eye-tracking research on teacher professional vision: A meta-analytic review[J]. Educational Research Review, 2023: 100586.
[13]. Lai H, Chen H, Wu S.Different Contextual Window Sizes Based RNNs for Multimodal Emotion Detection in Interactive Conversations[J].IEEE Access, 2020, 8:119516-119526.
[14]. Marks H M. Student engagement in instructional activity: Patterns in the elementary, middle, and high school years[J]. American educational research journal, 2000, 37(1): 153-184.
[15]. Newell A. Précis of unified theories of cognition[J]. Behavioral and Brain Sciences, 1992, 15(3): 425-437.
[16]. Pace C R. Measuring the quality of student effort[J]. Current issues in higher education, 1980, 2(1): 10-16.
[17]. Perez-Martin J, Bustos B,JorgePérez.Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding[C]//WACV 2021.2021.
[18]. Philp J, Duchesne S. Exploring engagement in tasks in the language classroom[J]. Annual Review of Applied Linguistics, 2016, 36: 50-72.
[19]. Radford A, Kim J W, Hallacy C, et al.Learning Transferable Visual Models from Natural Language Supervision[J]. 2021.
[20]. Reeve J. A self-determination theory perspective on student engagement[M]//Handbook of research on student engagement. Boston, MA: Springer US, 2012: 149-172.
[21]. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[22]. Xu W, Yu J, Miao Z, et al.Deep Reinforcement Polishing Network for Video Captioning[J].IEEE Transactions on Multimedia, 2020, PP(99):1-1.
[23]. Zhen L, Hu P, Wang X, et al.Deep Supervised Cross-Modal Retrieval[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).IEEE, 2020.
[24]. Zhou H, Sun Y, Wu W,etal.Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation.2021[2024-05-16].

Supported by the Graduate Scientific Research Foundation of Wenzhou University (Grant No. 3162024003016).