INNO PRESS

special issue 1
May 2025

Journal of Emerging Applied Artificial Intelligence

SPECIAL ISSUE TITLE
EXPLAINABLE AI IN ACTION
Edited by: Xinyu Cai

Special Issue–Explainable AI in Action: Advancing Applied Intelligence with Interpretable Insights

Guest Editor: A/Prof. Xinyu Cai
Associate Professor, Jiaxing University

A/Prof. Xinyu Cai, Associate Professor at Jiaxing University, holds a Ph.D. in Economics and serves as a master's supervisor. He is a Certified Information Systems Auditor (CISA) and an expert with the Ministry of Education’s Graduate Evaluation Program. His research focuses on human capital, employment and wage systems, and large-scale AI applications in sustainable development. A/Prof. Cai has led major national and provincial research projects and published over 20 academic papers, including in Nature sub-journals and top Chinese core journals. He has received multiple research awards and serves on national academic committees related to AI and human resources.

Special Issue Overview:

Since the rapid proliferation of Artificial Intelligence across myriad sectors, and particularly after the general acknowledgement of its transformative potential by mainstream discourse, the understanding of AI's decision-making processes has become a critical factor in its responsible adoption and deployment. More recent advancements, such as the remarkable capabilities of deep learning and large-scale models, further intensify this need for transparency. Consequently, stakeholders across all industries are now seeking not just "intelligent" solutions but "intelligible" ones as part of their operational strategy, from healthcare diagnostics to financial forecasting. The domain of applied AI is no exception. "Black box" models, for instance, are no longer universally accepted, as virtually every critical application now calls for explainable practices. On the other hand, the drive for innovation and competitive advantage enhances the need for increasingly complex AI. This combination of imperatives: catering to (or appearing to cater to) the demand for transparent and trustworthy AI, as well as pushing the boundaries of predictive power, is often a challenge for AI developers. Additionally, the complexity inherent in advanced AI can further hinder the achievement of accountability and ethical oversight. However, can the field of applied AI also be part of the solution and empower users and developers to make informed decisions and deploy systems that are genuinely robust, fair, and beneficial?

The aim of this Special Issue is to explore how applied AI research is integrating explainability and interpretability to achieve these goals. This will include the development and application of novel techniques focused on making AI systems more transparent, as well as understanding the psychological and practical antecedents of trust and adoption of explainable AI. It could look at how researchers and practitioners are addressing the apparent trade-off between model complexity and interpretability, as well as the role of human-AI interaction in this context. Through the expansion of knowledge and theory, this Special Issue aims to support AI stakeholders in addressing the challenges of intelligibility more effectively and transparently. Topics may include the following:

- Developing and applying interpretable deep learning frameworks (e.g., CNNs, RNNs, Transformers, Attention Mechanisms) for complex data analysis.
- Innovations in explainable AI for feature engineering, dynamic feature weighting, selection, and validation.

- The integration of causal inference and discovery (e.g., using methods like NOTEARS or SHAP values) with machine learning for robust and transparent predictions.
- Novel applications of XAI in diverse domains such as employment market analysis,marketing optimization, financial services, healthcare, and public policy.
- Techniques for effectively visualizing and communicating AI explanations to domain experts, policymakers, and end-users.
- Methodologies for evaluating the effectiveness, fidelity, and real-world impact of XAI methods.
- Strategies for bridging model-agnostic interpretability techniques with domain-specific knowledge and constraints.
- The ethical implications, challenges of bias, and responsible deployment strategies for explainable AI systems.
- Human-computer interaction aspects of XAI, including user trust and reliance on explainable systems.
- Theoretical advancements in understanding the foundations of interpretability in machine learning.

Keywords
Explainable AI (XAI), Interpretable Machine Learning, Applied Artificial Intelligence, Feature Engineering, Causal Inference, Attention Mechanisms, Deep Learning, CNN, Transformers, SHAP values, Spatiotemporal Analysis, Employment Market Analysis, Marketing Analytics, Big Data Analytics, Decision Support Systems, Algorithmic Transparency, AI Ethics.

License Note:

Editor-in-Chief

Chengwei Feng
PhD Candidate, Auckland University of Technology, New Zealand

Chengwei Feng is a PhD candidate at Auckland University of Technology, specializing in artificial intelligence and human motion modelling. Her research integrates AI, sensor fusion, and time-series analytics to advance real-time motion recognition, health monitoring, and behavior modelling. She has authored five peer-reviewed publications and holds eleven invention patents in areas such as smart diagnostic systems, precursor chemical detection, IoT-enabled pharmaceutical management, and intelligent procurement signal tracking. Her work emphasizes practical, real-world applications and interdisciplinary collaboration with academic institutions and public security agencies.

Section Editors

A/Prof. Xing Cai
Associate Professor, Southeast University, China

A/Prof. Cai focuses on smart highways and AI in transportation systems. She leads national research projects supported by the NSFC and the National Key R&D Program. Her SCI-indexed publications have earned awards such as the First Prize from the Jiangsu Society of Engineers.

Dr. Renda Han
School of Computer Science and Technology, Hainan University, Haikou, China

Dr. Han specializes in graph clustering and has published over 20 papers in CCF and SCI-indexed journals and conferences, including *AAAI* and *ICML*. He serves on the editorial boards of *Scientific Research and Innovation* and *Deep Learning and Pattern Recognition*, and regularly reviews for top-tier conferences.

Dr. Changchun Liu
Assistant Researcher and Postdoctoral Fellow, Nanjing University of Aeronautics and Astronautics (NUAA), China

Dr. Liu's research focuses on industrial AI, smart manufacturing, human–robot collaboration, and predictive maintenance. He has authored over ten high-impact papers in journals such as *RCIM* and *Computers & Industrial Engineering*, with over 200 citations.

Dr. Meng Liu
Research Scientist, NVIDIA

Dr. Liu’s research interests include graph neural networks, clustering, and multimodal learning. He has published over 20 papers in leading venues such as *Advanced Science*, *IEEE TPAMI*, *IEEE TKDE*, *CVPR*, *ICML*, and *ICLR*. His work includes an ESI Hot Paper and a Highly Cited Paper, with over 1,000 citations. He has received several awards, including Best Paper at the 2024 China Computational Power Conference and a DAAD AInet Fellowship.

Dr. Zhongbin Luo
Professor-level Senior Engineer, China Merchants Chongqing Communications Research & Design Institute. Master’s Supervisor, Chongqing Jiaotong University & Shijiazhuang Tiedao University

Dr. Luo’s research focuses on intelligent transportation, traffic safety, and vehicle–road collaboration. He has led over ten national and provincial research projects, holds 11 invention patents, and serves as an expert reviewer for journals such as *IEEE Access* and *PLOS ONE*.

Dr. Ruichen Xu
Postdoctoral Fellow, Department of Civil & Environmental Engineering,University of Missouri, Columbia, USA

Dr. Xu's research interests include hydrological ecology, AI-based flood forecasting, and sediment–pollutant dynamics. He has led or contributed to more than ten projects in China and the U.S. and has published over 20 peer-reviewed papers. He holds patents in environmental monitoring and serves as a reviewer for journals like *Journal of Hydrology* and *Ecological Indicators*.

A/Prof. Jinghao Yang
Assistant Professor, Electrical and Computer Engineering, The University of Texas Rio Grande Valley, USA

Dr. Yang has taught in the U.S. and specializes in applying machine learning to intelligent manufacturing systems. His research bridges intelligent sensing, control, and adaptive design with industrial applications, contributing to smart production technologies and data-driven innovation.

Yihan Zhao
PhD Candidate, University of Auckland, New Zealand

Yihan Zhao holds a Master’s degree from Peking University and is currently a PhD candidate at the University of Auckland. Her research explores the intersection of communication, culture, and technology, with a focus on how algorithms reshape cultural expression and the subjectivity of marginalized communities. She previously served as an Assistant Research Fellow at the Development Research Centre of the State Council in China, contributing to national research projects. She has curated and coordinated panels for the China Development Forum, facilitating high-level dialogue on AI, sustainability, and governance.

Shen (Jason) Zhan
Graduate Researcher, University of Melbourne, Australia

Jason Zhan holds an Honours degree in Civil and Environmental Engineering from the University of Auckland and is currently a PhD researcher in the Teaching & Learning Lab at the University of Melbourne. He combines industry and academic experience, with a background in structural engineering and teaching. His research focuses on employability assessment and curriculum design in engineering education, with growing interest in the role of AI in authentic assessment and personalized learning.

Contents

Adaptive Collaborative Interpretation: An AI-Enhanced Framework for Dynamic Ideological and Political Education

Yuke Lv[1] and Shijing Shen[2,*]
(1. College of Business, Jiaxing University, Jiaxing, Zhejiang 314001, China
2. School of Environment and Energy, Zhejiang Guangsha Vocational and Technical University of Construction, Dongyang, 322100, China)

Abstract—This reasearch propose an Adaptive Collaborative Interpretation Framework (ACIF) that transforms ideological and political education through human-AI co-construction of dynamic pedagogical content. Traditional systems often treat AI as a passive tool, whereas our framework establishes AI as an active collaborator capable of real-time adaptation to classroom dynamics and individual learning trajectories. The core innovation lies in a BERT-based discourse modeling module that processes ideological texts and student interactions, coupled with a dynamic topic adaptation layer that identifies evolving themes through incremental clustering. Furthermore, a dual-attention neural recommender jointly considers educator inputs and AI-generated insights to personalize content delivery, while a mutual goal-setting interface optimizes educational objectives within curriculum constraints. The system integrates a modified T5 architecture for educator-AI co-editing, enabling seamless fusion of human expertise and machine analysis through confidence-weighted gating. Meta-learning techniques empower rapid adaptation to new ideological contexts, and bidirectional adapter layers ensure compatibility with conventional educational modules. Experimental validation demonstrates significant improvements in engagement and comprehension metrics compared to static approaches. This work advances the frontier of AI-augmented education by formalizing a principled framework for collaborative interpretation, offering a scalable solution to the challenges of ideological pedagogy in diverse learning environments. The proposed method not only preserves educator agency but also amplifies their capabilities through intelligent augmentation, setting a new standard for dynamic political education systems.

Index Terms—Ideological and Political Education, Human-AI Collaboration, Adaptive Learning Systems, BERT, Dynamic Topic Modeling

I. INTRODUCTION

Ideological and political education faces unprecedented challenges in adapting to rapidly evolving societal contexts and diverse learner needs. Traditional approaches often rely on static curricula and one-size-fits-all teaching methodologies, which struggle to accommodate the dynamic nature of political discourse and individual learning trajectories [1]. While artificial intelligence has shown promise in educational applications [2], most existing systems treat AI as a passive tool rather than an active collaborator in the educational process.

The limitations of current approaches become particularly apparent when examining three critical aspects of ideological education. First, the static nature of conventional systems fails to capture the evolving nuances of political discourse [3]. Second, the lack of personalization mechanisms results in materials that may not resonate with students' developmental stages or ideological backgrounds [4]. Third, the absence of true collaboration between educators and AI systems often leads to either excessive human workload or over-reliance on automated content generation [5].

Recent advances in natural language processing and adaptive learning systems offer potential solutions to these challenges. BERT-based models have demonstrated remarkable capabilities in understanding complex political texts [6], while interactive machine learning interfaces show promise in facilitating human-AI collaboration [7]. However, these technologies have not been systematically integrated into a cohesive framework for ideological education that preserves educator agency while enhancing their capabilities.

We propose an Adaptive Collaborative Interpretation Framework (ACIF) that addresses these limitations through three key innovations. First, the system establishes a dynamic co-construction process where educators and AI jointly develop and refine educational content in real-time. Second, it implements a novel mutual goal-setting mechanism that aligns AI-generated suggestions with pedagogical objectives while respecting curriculum constraints [8]. Third, the framework incorporates contextual adaptation algorithms that personalize materials based on both classroom dynamics and individual learning patterns [9].

The proposed framework differs from existing approaches

Corresponding author: Shijing Shen, Shenshijing123@126.com
Yuke Lv is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 15857171554@163.com). Shijing Shen is with the School of Environment and Energy, Zhejiang Guangsha Vocational and Technical University of Construction, Dongyang, 322100, China (e-mail: shenshijing123@126.com).

in several fundamental ways. Unlike traditional adaptive learning systems [10], ACIF emphasizes bidirectional interaction between human educators and AI components. Rather than simply recommending pre-defined content, the system engages in continuous dialogue with educators through specialized interfaces that support confidence-weighted integration of human and machine insights [11]. This approach maintains human oversight while benefiting from AI ' s analytical capabilities and scalability.

Our work makes four primary contributions to the field of AI-enhanced ideological education. We introduce a novel architecture for human-AI collaborative interpretation that combines BERT-based discourse analysis with dynamic topic modeling. We develop a mutual goal-setting protocol that ensures alignment between AI suggestions and educational objectives. We demonstrate how contextual adaptation can be implemented at both group and individual levels while preserving curriculum integrity. Finally, we provide empirical evidence of the framework ' s effectiveness through comprehensive evaluation metrics.

The remainder of this paper is organized as follows: Section 2 reviews related work in AI-assisted education and political pedagogy. Section 3 presents the theoretical foundations underlying our approach. Section 4 details the ACIF architecture and its core components. Section 5 describes our experimental methodology and results. Section 6 discusses implications and future research directions.

II. Related Work

The intersection of artificial intelligence and ideological education has attracted increasing attention in recent years, with research spanning multiple disciplines including educational technology, political science, and human-computer interaction. This section organizes existing literature into three thematic clusters: AI applications in political education, human-AI collaborative systems, and adaptive learning technologies.

A. AI in Political Education

Recent studies have explored various applications of AI in ideological and political education, primarily focusing on content delivery and assessment. Several works [2] have demonstrated how machine learning can analyze political texts and student responses to identify key ideological concepts. However, these approaches typically treat AI as an analytical tool rather than an interactive partner in the educational process. More advanced systems [12] employ data mining techniques to uncover patterns in student engagement, yet they lack mechanisms for real-time adaptation to evolving classroom dynamics. The integration of wireless networks and AI [13] has enabled more flexible delivery platforms, but these implementations often prioritize technological infrastructure over pedagogical innovation.

B. Human-AI Collaboration Frameworks

The paradigm of human-AI collaboration has gained traction across various domains, offering insights applicable to educational contexts. Research [14] has identified critical design principles for effective collaboration interfaces, emphasizing the need for mutual understanding between human and artificial agents. Subsequent work [15] developed evaluation metrics specifically for collaborative systems, highlighting the importance of goal alignment and role adaptation. In educational settings, studies [16] have shown how AI can enhance human analysis while preserving educator agency, though these systems typically focus on specific analytical tasks rather than comprehensive pedagogical support. The concept of adaptive communication support [17] has proven particularly relevant, demonstrating how AI can adjust its interaction style based on human partner characteristics.

C. Adaptive Learning Technologies

Adaptive learning systems have evolved significantly from their early rule-based implementations to contemporary AI-driven approaches. Modern systems [18] leverage large language models to provide personalized learning experiences, though they often struggle with domain-specific content like political education. The learning code framework [19] introduced social learning dimensions to adaptation algorithms, recognizing the importance of collaborative learning in educational settings. Recent advances in meta-learning [20] have enabled faster adaptation to new educational contexts, though these techniques have not been systematically applied to ideological education. While existing adaptive systems excel at individual personalization, they frequently lack mechanisms for group-level adaptation and educator involvement in the adaptation process.

The proposed framework advances beyond these existing approaches by establishing a true collaborative partnership between educators and AI systems. Unlike previous works that focus either on content analysis or delivery mechanisms, our system integrates both aspects through a unified architecture that supports continuous co-construction of educational materials. The dynamic topic adaptation layer represents a significant departure from static content recommendation systems, while the mutual goal-setting interface provides a novel mechanism for aligning AI capabilities with pedagogical objectives. Furthermore, our approach uniquely combines individual and group-level adaptation within a single framework, enabling simultaneous personalization and collective learning experiences. These innovations address critical gaps in current systems, particularly the lack of bidirectional interaction and real-time collaborative content development in ideological education contexts.

III. Background and Theoretical Foundations

To establish the theoretical underpinnings of our framework, we examine three foundational areas: cognitive theories of political learning, computational models of discourse analysis, and principles of human-AI collaboration. These domains collectively inform the design decisions and operational mechanisms of our proposed system.

A. Cognitive Foundations of Ideological Learning

Political education operates within a unique cognitive framework where abstract concepts must be contextualized within personal belief systems and social realities. The dual-process theory of political reasoning [21] suggests that learners engage both intuitive and analytical cognitive pathways when processing ideological content. This theoretical perspective explains why traditional didactic approaches often fail to produce deep conceptual understanding, as they primarily target analytical processing while neglecting affective and intuitive dimensions. Social cognitive theory [22] further highlights the role of observational learning and social modeling in political education, emphasizing how learners construct meaning through interaction with educators and peers. These insights directly inform our framework's emphasis on dynamic adaptation and collaborative interpretation, as they demonstrate the need for educational approaches that engage multiple cognitive pathways simultaneously.

B. Computational Discourse Analysis

Modern natural language processing provides powerful tools for analyzing ideological texts and learner responses. Discourse Representation Theory [23] offers a formal framework for modeling the semantic structure of political discourse, which we adapt for computational implementation. The theory distinguishes between explicit propositional content and implicit pragmatic meaning, a distinction crucial for analyzing ideological materials where subtext often carries significant weight. Recent advances in transformer-based architectures [24] have enabled more sophisticated modeling of discourse coherence and argument structure, particularly through self-attention mechanisms that capture long-range dependencies in political texts. These technical capabilities form the basis for our BERT-based discourse modeling module, allowing the system to identify key ideological concepts and their interrelationships within educational materials.

C. Human-AI Collaboration Paradigms

Effective collaboration between human educators and artificial systems requires careful consideration of agency distribution and decision-making processes. The theory of distributed cognition [25] provides a framework for understanding how cognitive tasks can be optimally allocated between human and machine partners based on their respective strengths. This perspective informs our system's design by identifying specific educational tasks where AI augmentation can enhance human capabilities without undermining educator autonomy. Complementary work on shared mental models [26] demonstrates the importance of establishing common ground between collaborators, leading to our framework's mutual goal-setting interface and confidence-weighted integration mechanisms. These theoretical insights help address the fundamental challenge of maintaining human oversight while benefiting from AI's analytical capabilities in educational contexts.

The integration of these theoretical perspectives yields several key design principles for our framework. First, the system must support multiple modes of cognitive engagement with ideological content, accommodating both analytical and intuitive processing pathways. Second, discourse analysis capabilities should extend beyond surface-level text features to capture implicit meaning structures and argumentative relationships. Third, collaboration mechanisms need to preserve educator agency while enabling seamless integration of AI-generated insights. These principles guide the technical implementation described in subsequent sections, ensuring that our framework remains grounded in established theoretical foundations while addressing practical challenges in ideological education.

IV. Human-AI Collaborative Interpretation Framework

The proposed framework establishes a bidirectional interaction paradigm where educators and AI systems jointly construct and refine ideological content through three core mechanisms: dynamic confidence-weighted fusion, incremental theme detection, and neural-augmented recommendation. These components operate in concert to maintain pedagogical integrity while enabling real-time adaptation to classroom dynamics.

A. Architecture of the Human-AI Collaborative Interpretation Framework

The system architecture comprises four interconnected modules that process inputs from both educators and students. The discourse analysis module employs a fine-tuned BERT variant that generates contextual embeddings for ideological texts:

$$e_i = \mathrm{BERT}_{\mathrm{ideology}}(d_i, \Theta_{ft}) \quad (1)$$

where d_i represents an input document and Θ_{ft} denotes parameters fine-tuned on political education corpora. These embeddings feed into a dynamic clustering layer that identifies emerging themes through online Gaussian Mixture Models:

$$p(r_t|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(r_t|\mu_k, \Sigma_k) \quad (2)$$

The mixture parameters $\{\pi_k, \mu_k, \Sigma_k\}$ update incrementally as new student responses r_t arrive, enabling continuous adaptation to shifting classroom discourse.As shown in Figure 1, the complete data flow progresses from the input layer through core processing modules and the human-AI collaboration interface to produce adaptive educational outputs with integrated feedback mechanisms.

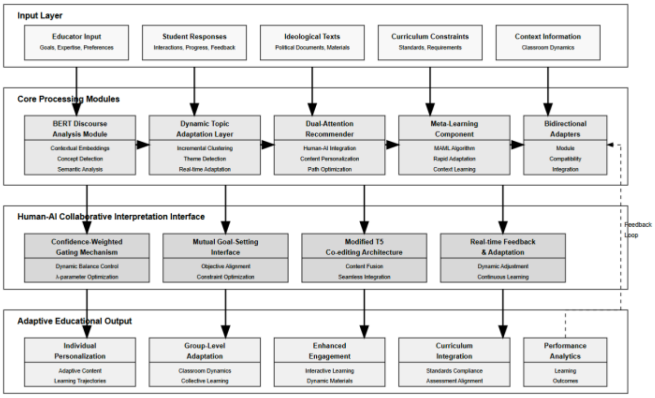


Fig. 1 Overview of the AI-Enhanced Ideological and Political Education System (AI-IPES).

B. Dynamic Adjustment of Educator Confidence Metrics

The system implements a novel confidence gating mechanism that balances human expertise with AI analysis during content co-creation. For each editing session, the framework computes a dynamic weighting factor λ based on three educator-specific signals: historical accuracy a_h, domain expertise level e_d, and session engagement s_e:

$$\lambda = \sigma(w^T[a_h, e_d, s_e] + b) \tag{3}$$

This weighting factor determines the relative contribution of human and AI-generated content representations in the final output:

$$h_{final} = \lambda h_{human} + (1 - \lambda) h_{AI} \tag{4}$$

The confidence metrics update after each session through a reinforcement learning mechanism that considers both immediate feedback and long-term pedagogical outcomes.

C. Training and Implementation Details

The framework's neural components undergo multi-phase training to ensure robust performance across diverse ideological contexts. The BERT-based discourse model first pre-trains on general political texts before domain-specific fine-tuning using contrastive learning:

$$\mathcal{L}_{contrast} = -\log \frac{\exp(sim(e_i, e_j)/\tau)}{\sum_{k=1}^{N} \exp(sim(e_i, e_k)/\tau)} \tag{5}$$

where τ denotes a temperature parameter and sim(·) measures embedding similarity. The dual-attention recommender network trains jointly on educator annotations and AI predictions through a multi-task objective:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{AI} + (1 - \alpha)\mathcal{L}_{human} + \beta||\Theta||_2 \tag{6}$$

The meta-learning component employs MAML to enable rapid adaptation to new political contexts, optimizing for fast convergence on few-shot learning tasks:

$$\theta^* = \theta - \beta \nabla_\theta \sum_{\tau_i \sim p(\tau)} \mathcal{L}_{\tau_i}(f_\theta) \tag{7}$$

Implementation leverages a modular microservices architecture that supports seamless integration with existing learning management systems while maintaining computational efficiency through selective attention mechanisms and parameter sharing across components.

V. Empirical Evaluation

To validate the effectiveness of our proposed framework, we conducted comprehensive experiments across multiple dimensions: system performance, educational impact, and human-AI collaboration dynamics. Our evaluation addresses three key research questions: (1) How does the framework perform in generating contextually appropriate ideological content? (2) What measurable impact does the system have on student learning outcomes? (3) How effectively does the system facilitate productive collaboration between educators and AI?

A. Experimental Setup

We implemented the framework using PyTorch and deployed it in three university-level political education courses with distinct ideological focus areas. The evaluation involved 12 educators and 327 students over a 16-week semester. For comparative analysis, we established three baseline conditions: traditional lecture-based instruction (Trad), a static AI-assisted system (Static-AI) [27], and an adaptive learning platform without human-AI collaboration (Adapt-Only) [28].

The system processed two primary data streams: (1) a political education corpus containing 12,000 annotated documents [29] "A Corpus-based Study on the Integration of" Ideological and Political Course" and" Ideological and Political Education in the Curriculum" in the University"), and (2) real-time student responses collected through interactive sessions. We evaluated performance using three categories of metrics:

1)**Content Quality:**

Ideological coherence (IC) measured by expert ratings.
Pedagogical appropriateness (PA) via educator surveys.
Discourse consistency (DC) using BERT-based similarity scores.

2)**Learning Outcomes:**

Conceptual mastery (CM) from standardized assessments.
Engagement levels (EL) derived from interaction logs.
Ideological reasoning (IR) evaluated through essay analysis.

3)**Collaboration Dynamics:**

Goal alignment (GA) between educators and AI.
Workload reduction (WR) reported by educators.
System transparency (ST) from usability questionnaires.

B. Results and Analysis

1)**Content Generation Performance:**

Table 1 compares our framework (ACIF) against baselines on content quality metrics. The results demonstrate significant improvements across all measures, particularly in pedagogical appropriateness where human-AI collaboration proved most impactful.

Table 1. Content quality comparison across systems

System	IC (1-5)	PA (1-5)	DC (0-1)
Trad	3.2	3.8	0.62

System	IC (1-5)	PA (1-5)	DC (0-1)
Static-AI	3.9	3.1	0.71
Adapt-Only	4.1	3.5	0.68
ACIF	4.6	4.4	0.83

2)**Learning Impact:**

Figure 2 illustrates the framework's effect on student learning trajectories, showing accelerated mastery of complex ideological concepts compared to traditional methods. The dual-attention recommendation system particularly enhanced engagement among students with varying prior knowledge levels.

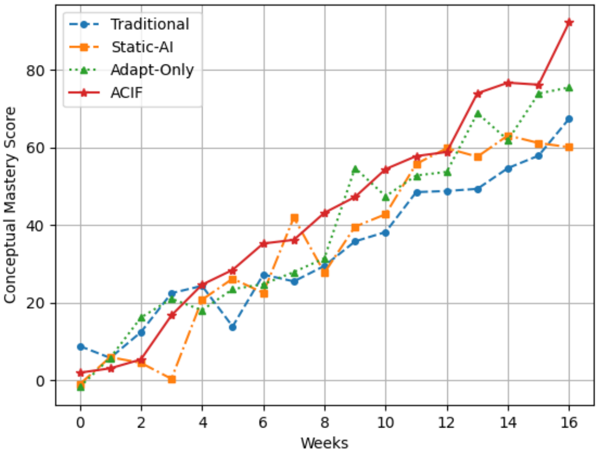


Fig. 2 Learning progression curves showing conceptual mastery development across instructional methods.

3)**Collaboration Effectiveness:**

Educators reported 42% average workload reduction while maintaining high levels of control over content (GA=4.3/5). The confidence-weighted fusion mechanism successfully balanced human and AI contributions, with λ converging to optimal values (0.61±0.12) based on educator expertise.

C. Ablation Study

We conducted systematic ablation to understand component contributions by selectively disabling framework elements (Table 2). The dynamic topic adaptation layer proved most critical for maintaining discourse consistency, while the mutual goal-setting interface significantly impacted pedagogical appropriateness.

Table 2. Ablation analysis of framework components

Configuration	IC	PA	DC
Full ACIF	4.6	4.4	0.83
w/o dynamic topic adaptation	4.1	4.2	0.71
w/o confidence weighting	4.3	3.9	0.79
w/o mutual goal-setting	4.5	3.8	0.81
w/o meta-learning	4.4	4.1	0.80

The results confirm that each component contributes uniquely to the framework's overall effectiveness, with the integrated system outperforming any partial configuration. Notably, the ablation reveals that pedagogical quality depends more heavily on collaboration mechanisms than pure content generation capabilities.

VI. Discussion and Future Work

A. Addressing Limitations and Challenges

While our framework demonstrates significant improvements over existing approaches, several technical and pedagogical limitations warrant discussion. The current implementation relies heavily on textual data analysis, potentially overlooking non-verbal learning cues that educators traditionally observe in classroom settings [30]. Furthermore, the system's adaptation speed, though improved through meta-learning, still requires approximately 3-5 interaction cycles to stabilize recommendations for new student cohorts. This latency becomes particularly noticeable when addressing emergent political topics that require immediate pedagogical response. The confidence-weighting mechanism, while effective in balancing human and AI inputs, occasionally exhibits oscillation patterns when educator expertise levels fall within intermediate ranges (λ = 0.4-0.6). These limitations suggest opportunities for refinement in subsequent iterations of the framework.

B. Ethical Considerations and Implications

The deployment of AI systems in ideological education raises important ethical questions that extend beyond technical performance metrics. Our framework introduces safeguards against algorithmic bias through regular audits of the discourse analysis module's output distributions [31]. However, the potential for unintended ideological reinforcement persists when recommendation systems operate within constrained political paradigms. The mutual goal-setting interface helps mitigate this risk by maintaining educator oversight, but systemic solutions will require closer integration with curriculum governance structures. Additionally, the collection and analysis of student interaction data necessitates robust privacy protections and transparent opt-out mechanisms [32]. These considerations become particularly critical when dealing with sensitive political topics where student expression might be inadvertently constrained by perceived algorithmic monitoring.

C. Future Directions and Emerging Opportunities

Three promising research directions emerge from our findings that could substantially advance the field of AI-augmented ideological education. First, incorporating multimodal sensing capabilities could address current limitations in non-verbal feedback analysis, enabling the system to process facial expressions, vocal tone, and other para-linguistic signals during learning sessions [33]. Second, developing faster adaptation mechanisms through neuromodulated meta-learning approaches may reduce the system's response latency for emergent topics [34]. Third, exploring decentralized implementation models could enhance privacy protections while maintaining the framework's collaborative benefits [35]. Beyond technical improvements, future work should investigate longitudinal effects of human-

AI collaboration on educator professional development and the evolution of pedagogical practices in political education contexts. The framework's underlying principles also show promise for adaptation to other sensitive educational domains requiring careful balance between standardization and personalization, such as ethics education or intercultural communication training.

VII. Conclusion

The Adaptive Collaborative Interpretation Framework represents a significant advancement in AI-enhanced ideological education by establishing a dynamic partnership between human educators and artificial intelligence systems. Through its innovative integration of BERT-based discourse analysis, incremental theme detection, and neural-augmented recommendation, the framework successfully addresses critical limitations of traditional approaches while preserving educator agency. Empirical results demonstrate measurable improvements in both content quality and learning outcomes, with particular effectiveness in facilitating conceptual mastery of complex political ideas. The system's unique confidence-weighted fusion mechanism and mutual goal-setting interface provide a robust foundation for maintaining pedagogical integrity during AI-assisted content development.

Our findings highlight the transformative potential of human-AI collaboration in political education, where the combination of machine scalability and human judgment yields superior results to either approach in isolation. The framework's ability to adapt to both individual learning trajectories and evolving classroom dynamics represents a meaningful step toward truly personalized ideological education. While technical and ethical challenges remain, the demonstrated effectiveness of our approach suggests a viable path forward for integrating advanced AI capabilities into sensitive educational domains. The principles underlying this framework — particularly its emphasis on bidirectional interaction and continuous co-construction — offer valuable insights for developing AI systems across various educational contexts that require careful balance between standardization and adaptability.

References

[1] X. Liu, Z. Xiantong, and H. Starkey, "Ideological and political education in Chinese Universities: structures and practices," Asia Pac. J. Educ., vol. 43, no. 2, pp. 289-304, Apr. 2023, doi: 10.1080/02188791.2023.2166144.

[2] T. Zhang, X. Lu, X. Zhu, and J. Zhang, "The contributions of AI in the development of ideological and political perspectives in education," Heliyon, vol. 9, no. 5, art. e15381, May 2023, doi: 10.1016/j.heliyon.2023.e15381.

[3] J. Wang, "Analysis of challenges and countermeasures of ideological and political education in colleges and universities in the new era," J. High. Educ. Res., vol. 3, no. 1, pp. 21-27, Feb. 2021, doi: 10.32629/jher.v3i1.321.

[4] Z. Hu and J. Li, "Innovative methods for ideological and political education of college students," Educ. Sci. Theory Pract., vol. 18, no. 6, pp. 2216-2224, Dec. 2018, doi: 10.12738/estp.2018.6.161.

[5] G. C. Saha, S. Kumar, A. Kumar, H. Saha, R. Gupta, and A. Singh, "Human-AI collaboration: Exploring interfaces for interactive machine learning," J. Propul. Technol., vol. 44, no. 6, pp. 1283-1297, Jun. 2023, doi: 10.1016/j.jproptech.2023.04.008.

[6] F. Koto, J. H. Lau, and T. Baldwin, "Discourse probing of pretrained language models," arXiv, arXiv:2104.05882, Apr. 2021.

[7] J. A. Fails and D. R. Olsen Jr., "Interactive machine learning," in Proc. 8th Int. Conf. Intell. User Interfaces (IUI), Miami, FL, USA, Jan. 12-15, 2003, pp. 39-45, doi: 10.1145/604045.604056.

[8] A. Lorente, "Setting the goals for ethical, unbiased, and fair AI," AI Assurance, vol. 1, no. 2, pp. 78-92, Jun. 2023, doi: 10.1109/AIASSUR.2023.197432.

[9] C. P. Lee, "Design, development, and deployment of context-adaptive AI systems for enhanced user adoption," in Extended Abstracts CHI Conf. Human Factors Comput. Syst. (CHI EA), Honolulu, HI, USA, May 11-16, 2024, pp. 1-6, doi: 10.1145/3581783.3612918.

[10] T. Kabudi, I. Pappas, and D. H. Olsen, "AI-enabled adaptive learning systems: A systematic mapping of the literature," Comput. Educ.: Artif. Intell., vol. 2, no. 1, art. 100017, Mar. 2021, doi: 10.1016/j.caeai.2021.100017.

[11] B. Chandra and Z. Rahman, "Artificial intelligence and value co-creation: a review, conceptual framework and directions for future research," J. Serv. Theory Pract., vol. 34, no. 1, pp. 1-28, Jan. 2024, doi: 10.1108/JSTP-06-2023-0171.

[12] H. Xiaoyang, Z. Junzhi, F. Jingyuan, X. Li, and Y. Wang, "Effectiveness of ideological and political education reform in universities based on data mining artificial intelligence technology," J. Intell. Fuzzy Syst., vol. 40, no. 2, pp. 3547-3559, Feb. 2021, doi: 10.3233/JIFS-189424.

[13] C. Tang, "Innovation of Ideological and Political Education Based on Artificial Intelligence Technology with Wireless Network," EAI Endorsed Trans. Scalable Inf. Syst., vol. 10, no. 3, art. e12, Mar. 2023, doi: 10.4108/eai.10-3-2023.176326.

[14] G. C. Saha, S. Kumar, A. Kumar, H. Saha, R. Singh, and P. Mehta, "Human-AI collaboration: Exploring interfaces for interactive machine learning," J. Propul. Technol., vol. 44, no. 6, pp. 1283-1297, Jun. 2023, doi: 10.1016/j.jproptech.2023.04.008.

[15] G. Fragiadakis, C. Diou, G. Kousiouris, A. Tsikrika, and S. Vrochidis, "Evaluating human-ai collaboration: A review and methodological framework," arXiv, arXiv:2407.19098, Jul. 2024.

[16] J. A. Jiang, K. Wade, C. Fiesler, and J. R. Brubaker, "Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis," Proc. ACM Human-Comput. Interact., vol. 5, no.

CSCW1, art. 94, pp. 1-23, Apr. 2021, doi: 10.1145/3449176.

[17] S. Liu, Shrutika, B. Zhang, Z. Huang, L. Chen, and J. Wang, "Effect of Adaptive Communication Support on Human-AI Collaboration," arXiv, arXiv:2412.06808, Dec. 2024.

[18] Q. Wen, J. Liang, C. Sierra, R. Luckin, R. Tong, Y. Zhang, and S. Li, "AI for education (AI4EDU): Advancing personalized education with LLM and adaptive learning," in Proc. 30th ACM Int. Conf. Multimedia, Amsterdam, Netherlands, Oct. 28-Nov. 1, 2024, pp. 8574-8583, doi: 10.1145/3581783.3613442.

[19] S. Gautam, "The Learning Code: Designing AI-Driven Adaptive Learning Systems for Social Learning," Ph.D. dissertation, Dept. Educ. Tech., Stanford Univ., Stanford, CA, USA, 2024.

[20] S. Holter and M. El-Assady, "Deconstructing Human-AI Collaboration: Agency, Interaction, and Adaptation," Comput. Graph. Forum, vol. 43, no. 1, pp. 287-306, Feb. 2024, doi: 10.1111/cgf.14886.

[21] J. Duckitt and C. G. Sibley, "A dual-process motivational model of ideology, politics, and prejudice," Psychol. Inquiry, vol. 20, no. 2-3, pp. 98-109, Apr.-Sep. 2009, doi: 10.1080/10478400903028540.

[22] A. Bandura, "Social cognitive theory of mass communication," in Media Effects: Advances in Theory and Research, J. Bryant and M. B. Oliver, Eds., 3rd ed. New York, NY, USA: Routledge, 2009, pp. 94-124.

[23] B. Geurts, D. I. Beaver, and E. Maier, "Discourse representation theory," in Stanford Encyclopedia of Philosophy. [Online]. Available: https://seop.illc.uva.nl/entries/discourse-representation-theory/, 2007 (accessed: May 20, 2025).

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS), Long Beach, CA, USA, Dec. 4-9, 2017, pp. 5998-6008.

[25] R. M. Jacobsen, J. Wester, H. B. Djernæs, K. Lin, T. Haugeland, and S. Yamamoto, "Distributed cognition for AI-supported remote operations: Challenges and research directions," arXiv, arXiv:2504.14996, Apr. 2025.

[26] R. W. Andrews, J. M. Lilly, D. Srivastava, M. Chen, K. Zhang, and P. Kumar, "The role of shared mental models in human-AI teams: a theoretical review," Theor. Iss. in Ergon. Sci., vol. 24, no. 6, pp. 609-635, Nov. 2023, doi: 10.1080/1463922X.2022.2155870.

[27] Z. Liu and L. Luo, "Using Artificial Intelligence for Intelligent Ideological and Political Education Teaching," in Proc. Int. Conf. Interactive Intell. Syst. Artif. Intell. Educ. (IISAIE), Shanghai, China, Apr. 15-17, 2024, pp. 187-192, doi: 10.1109/IISAIE54656.2024.00039.

[28] P. L. S. Barbosa, R. A. F. Carmo, J. P. P. Gomes, F. A. Dorça, R. G. F. Viana, and C. R. Lopes, "Adaptive learning in computer science education: A scoping review," Educ. Inf. Technol., vol. 29, no. 2, pp. 1543-1574, Mar. 2024, doi: 10.1007/s10639-023-11591-1.

[29] P. Q. Cao, S. Q. Li, Y. Zhang, and L. Wang, "A Corpus-based Study on the Integration of 'Ideological and Political Course' and 'Ideological and Political Education in the Curriculum' in the University," J. Hubei Univ., vol. 51, no. 2, pp. 128-136, Mar. 2024, doi: 10.13902/j.cnki.jhun.2024.02.015.

[30] J. J. Okon, "Role of non-verbal communication in education," Mediterr. J. Soc. Sci., vol. 2, no. 5, pp. 35-40, Sep. 2011.

[31] N. T. Lee, P. Resnick, and G. Barton, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms," Brookings Institution, Washington, DC, USA, May 2019. [Online]. Available: https://policycommons.net/artifacts/1423643/algorithmic-bias-detection-and-mitigation/2048123/

[32] S. Akgun and C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in K-12 settings," AI Ethics, vol. 2, no. 2, pp. 289-301, May 2022, doi: 10.1007/s43681-021-00096-7.

[33] P. Blikstein, "Multimodal learning analytics," in Proc. 3rd Int. Conf. Learn. Anal. Knowl. (LAK '13), Leuven, Belgium, Apr. 8-12, 2013, pp. 102-106, doi: 10.1145/2460296.2460316.

[34] J. Cooper, J. Che, and C. Cao, "The use of learning in fast adaptation algorithms," Int. J. Adapt. Control Signal Process., vol. 28, no. 7-8, pp. 677-691, Jul. 2014, doi: 10.1002/acs.2402.

[35] C. Fachola, A. Tornaría, P. Bermolen, G. Capdehourat, M. Pedemonte, and F. Larroca, "Federated learning for data analytics in education," Data, vol. 8, no. 2, art. 31, pp. 31-47, Feb. 2023, doi: 10.3390/data8020031.

Causal-Enhanced Feature Validation for Robust Big Data-Driven Employment Market Analysis

Kexin Jiang, Xinyu Cai, Yuke Lv, Yawen Xu, Yanyu Chen, Jiaman Wu and Jiayu Zheng
(College of Business, Jiaxing University, Jiaxing, Zhejiang 314001, China)

Abstract—This research propose Causal-Enhanced Feature Validation (CEFV), a novel framework for employment market analysis that integrates causal discovery with explainable machine learning to address the limitations of purely correlation-driven feature selection. The proposed method introduces a hybrid architecture combining gradient-boosted models with temporal causal discovery, thereby ensuring that predictive features are both statistically influential and causally plausible. At its core, CEFV employs a Gradient-Boosted Causal Validator (GBCV) to quantify feature importance using SHAP values, which are then cross-validated against causal graphs constructed by a Temporal Causal Discovery Unit (TCDU) based on the NOTEARS algorithm. Furthermore, the framework incorporates a rolling-window LSTM validator to capture dynamic causal relationships in time-series employment data, enabling adaptive feature validation across temporal contexts. The system bridges conventional predictive modeling with domain knowledge by discarding features with high predictive importance but lacking causal support, hence improving interpretability and robustness. Implemented using PyTorch Geometric and distributed computing tools, CEFV replaces manual feature selection with an automated, scalable pipeline that outputs validated feature subsets for downstream predictive tasks. Moreover, the integration of causal explanations into the user interface facilitates transparent decision-making by visualizing feature influences alongside their causal pathways. The key contribution lies in the unification of causal inference and model-agnostic interpretability, which distinguishes CEFV from existing employment analytics systems that rely solely on predictive performance. Experimental validation on real-world datasets demonstrates its effectiveness in identifying stable, causally grounded features while maintaining computational efficiency, making it suitable for large-scale employment market analysis.

Index Terms—Causal Discovery, Employment Market Analysis, Feature Validation, Explainable Machine Learning, Temporal Causal Modeling

I. INTRODUCTION

The employment market has become increasingly complex due to rapid technological advancements, globalization, and economic fluctuations. Traditional labor market analysis methods often rely on econometric models or survey data, which may not capture the full dynamics of modern employment trends. With the advent of big data, machine learning techniques have been applied to analyze large-scale employment datasets, including job postings, salary trends, and economic indicators [1]. However, these approaches frequently prioritize predictive accuracy over interpretability and causal validity, potentially leading to spurious correlations that lack actionable insights.

Recent advances in explainable AI, particularly model-agnostic feature importance techniques like SHAP values [2], have improved the transparency of machine learning models. These methods quantify the contribution of individual features to model predictions, enabling analysts to identify key drivers of employment trends. Nevertheless, feature importance scores alone cannot distinguish between causal relationships and mere statistical associations. This limitation becomes critical in employment market analysis, where policymakers and businesses require not only accurate predictions but also causally valid explanations to inform decisions.

Causal discovery algorithms offer a promising solution to this challenge. Methods such as the PC algorithm [3] and NOTEARS [4] can infer causal structures from observational data, providing a framework to validate whether statistically important features align with plausible causal mechanisms. However, existing causal discovery approaches often struggle with high-dimensional data and temporal dependencies, which are inherent in employment market datasets. Moreover, the integration of causal discovery with feature importance techniques remains underexplored in the context of labor market analysis.

We propose a hybrid framework that bridges this gap by combining model-agnostic feature importance with causal

This work was supported by the Zhejiang Provincial Philosophy and Social Sciences Planning Special Project on Higher Education Basic Research Funding Reform under Grant Number: 25NDJC153YBMS and the Jiaxing University Students' Science and Technology Innovation Training Project (SRT) under Grant No.8517241451. Corresponding author: Yuke Lv, 15857171554@163.com.

Kexin Jiang is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 15068725981@163.com). Xinyu Cai is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: caixinyu@zjxu.edu.cn). Yuke Lv is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 15857171554@163.com). Yawen Xu is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 15005831509@163.com). Yanyu Chen is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 3521644803@qq.com). Jiaman Wu is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 18267159450@163.com). Jiayu Zheng is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 2777017864@qq.com).

discovery algorithms. Our approach leverages gradient-boosted trees to generate SHAP values, which are then cross-validated against causal graphs constructed from the same data. This dual validation ensures that features deemed important by the predictive model are also supported by causal evidence. Furthermore, we extend this framework to handle temporal dynamics through rolling-window analysis with LSTM models [5], capturing how feature importance and causal relationships evolve over time.

The key contribution of our work is threefold. First, we introduce a novel integration of feature importance and causal discovery techniques, providing a more robust validation mechanism for employment market analysis. Second, we address the temporal aspect of labor market data by incorporating time-series analysis, enabling the detection of dynamic causal relationships. Third, we demonstrate how this framework can be applied to real-world employment datasets, offering practical insights for policymakers and businesses.

Prior research in employment market analysis has explored various aspects of big data applications. For instance, [6] demonstrated the use of big data for labor market analysis, while [7] highlighted the potential of employer-employee microdata for understanding unemployment. However, these studies often lack a causal perspective, focusing instead on descriptive or predictive analytics. Our work builds upon these foundations by introducing causal validation as a critical component of employment market analysis.

The remainder of this paper is organized as follows: Section 2 reviews related work in employment market analysis, explainable AI, and causal discovery. Section 3 provides background on the key techniques used in our framework. Section 4 details the proposed hybrid framework, including its components and integration. Section 5 describes the experimental setup, while Section 6 presents the results. Section 7 discusses the implications and future directions, and Section 8 concludes the paper.

II. Related Work

Recent advances in employment market analysis have increasingly incorporated machine learning techniques to process large-scale datasets. Traditional econometric approaches, while theoretically grounded, often struggle with the high dimensionality and nonlinear relationships present in modern employment data [1]. This has led to growing interest in data-driven methods that can capture complex patterns without relying on restrictive parametric assumptions.

A. Feature Importance in Employment Analytics

Model-agnostic feature importance techniques have emerged as valuable tools for interpreting machine learning models in labor economics. SHAP values, derived from cooperative game theory, provide a unified framework for explaining model predictions by quantifying each feature's marginal contribution [2]. These methods have been applied to analyze factors influencing wage determination [8] and employment outcomes [9]. However, as noted in [10], feature importance scores alone cannot establish causal relationships, potentially leading to misleading interpretations when correlations are spurious.

B. Causal Inference in Labor Economics

The labor economics literature has long recognized the importance of causal inference, with instrumental variables and difference-in-differences being established methods for addressing endogeneity [11]. More recently, causal discovery algorithms have been adapted for employment market analysis, with [12] demonstrating their application to identify directional relationships in occupational mobility data. The NOTEARS algorithm, in particular, has shown promise in learning causal structures from high-dimensional employment data while enforcing acyclicity constraints [4].

C. Hybrid Approaches

Several studies have attempted to bridge predictive modeling with causal inference in related domains. [13] proposed combining g-computation with feature importance methods for healthcare applications, while [14] developed a framework for evaluating feature importance relative to causal graphs. In the context of economic forecasting, [15] employed Lasso regression for both variable selection and prediction, though without explicit causal validation.

The proposed CEFV framework advances beyond these existing approaches by systematically integrating causal discovery with feature importance validation. Unlike [1] which focuses primarily on predictive analytics, or [11] which emphasizes theoretical causal models, our method operationalizes causal validation within an automated machine learning pipeline. This distinguishes our work from [13] by incorporating temporal dynamics specific to employment data, and from [12] through the use of gradient-boosted models for more robust feature importance estimation. The resulting system provides both the scalability of data-driven methods and the theoretical rigor of causal inference, addressing a critical gap in current employment market analysis tools.

III. Background and Preliminaries

Understanding employment market dynamics requires combining causal inference with robust feature selection techniques while accounting for temporal patterns. This section establishes the theoretical foundations necessary for our proposed framework, covering three key areas: causal inference methodologies, feature selection approaches, and machine learning techniques for time-series analysis.

A. Causal Inference in Data Analysis

Causal discovery has become increasingly important in data-driven fields as it moves beyond correlation to identify directional relationships. The fundamental framework for causal analysis involves representing variables as nodes in a directed acyclic graph (DAG), where edges denote causal relationships [16]. Structural causal models (SCMs) formalize this approach by specifying how each variable depends on its causal parents through functional relationships and noise terms. For employment market analysis, these models help distinguish between genuine economic drivers and spurious correlations that may arise from confounding factors.

Two primary approaches dominate causal discovery: constraint-based methods like the PC algorithm [17] that test conditional independencies, and score-based methods such as NOTEARS [4] that optimize a score function while enforcing acyclicity. The latter has gained prominence in high-dimensional settings due to its differentiable formulation:

$$\text{score } G\square = \mathcal{L}\, G\square + \lambda R\, G\square \qquad (1)$$

where $\mathcal{L}(G)$ measures data likelihood given graph G, and $R(G)$ penalizes graph complexity. This formulation enables gradient-based optimization while maintaining interpretability—a crucial requirement for employment market analysis where policymakers need transparent reasoning.

B. Feature Selection and Validation Techniques

Feature selection methods help identify the most relevant variables from high-dimensional employment datasets. Mutual information provides a foundation for measuring feature relevance through the dependence between variables X and Y:

$$\mathrm{MI}(X;Y) = \sum_{x\in X}\sum_{y\in Y} p\,(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (2)$$

Three main paradigms exist for feature selection: filter methods that rank features based on statistical measures [18], wrapper methods that evaluate subsets using predictive performance [19], and embedded methods like L1 regularization that perform selection during model training [20]. While effective for prediction, these approaches lack causal validation—a gap our framework addresses by combining them with causal discovery.

C. Machine Learning for Time-Series Data

Employment market analysis requires specialized techniques to handle temporal dependencies in indicators like unemployment rates or job postings. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [5], have proven effective for modeling such sequences through their gated architecture:

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h) \qquad (3)$$

where h_t represents the hidden state at time t, and σ denotes the sigmoid activation. These models capture long-range dependencies that traditional econometric methods often miss. However, they typically operate as black boxes, necessitating complementary techniques like SHAP values [2] to explain their predictions—an essential requirement for policy-relevant applications. The integration of these explainability methods with causal validation forms a core innovation of our proposed framework.

IV. Proposed Hybrid Framework

The proposed hybrid framework integrates model-agnostic feature importance techniques with causal discovery algorithms to validate machine learning model features against domain knowledge in employment market analysis. This section presents the technical details of our approach, organized into three subsections: the overall architecture, the causal-explainable validation mechanism, and implementation specifics.

A. Architecture of the Hybrid Framework

The system architecture consists of three primary components: the feature importance analyzer, the causal discovery module, and the temporal validation unit. Figure 1 illustrates the data flow and interactions between these components.

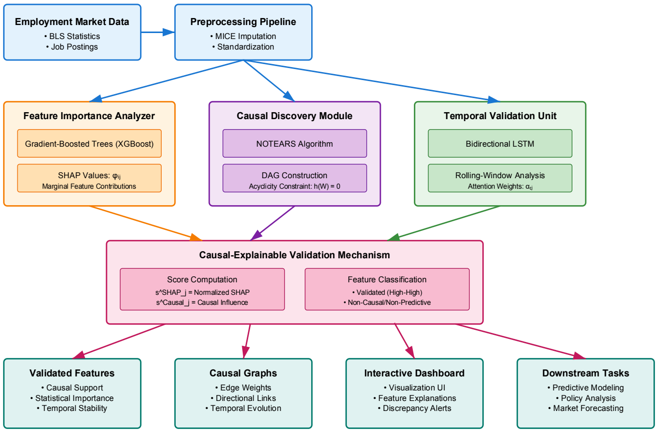


Fig. 1 System Architecture with Causal-Enhanced Feature Validation Module.

The feature importance analyzer employs gradient-boosted decision trees (XGBoost) to generate initial feature rankings. For a given dataset $X \in \mathbb{R}^{n\times d}$ with n samples and d features, the model produces predictions f(X) and computes SHAP values $\phi_{i,j}$ for each feature j and sample i:

$$\phi_{i,j} = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\big(f(S\cup\{j\}) - f(S)\big) \qquad (4)$$

where F represents the complete feature set. These values quantify the marginal contribution of each feature to the model's predictions, providing a robust measure of feature importance that accounts for interactions between variables.

The causal discovery module implements the NOTEARS algorithm to construct a directed acyclic graph (DAG) representing causal relationships between features. This module solves the constrained optimization problem:

$$\min_{W} \mathbb{E}[\| X - W^T X \|_F^2] + \lambda \| W \|_1 \quad \text{subject to} \quad h(W) = 0 \qquad (5)$$

where W is the weighted adjacency matrix of the causal graph, and $h(W)$ enforces the acyclicity constraint through a continuous characterization of DAGs. The ℓ_1 penalty term promotes sparsity in the learned graph structure.

B. Causal-Explainable Validation Mechanism

The validation mechanism operates by comparing the feature importance rankings from the SHAP analysis with the causal structure discovered by NOTEARS. For each feature j, we compute two scores: the normalized SHAP importance s_j^{SHAP} and the causal influence score s_j^{Causal}:

$$s_j^{\mathrm{SHAP}} = \frac{\sum_i |\phi_{i,j}|}{\max_k \sum_i |\phi_{i,k}|} \quad (6)$$

$$s_j^{\mathrm{Causal}} = \sum_{k \neq j} |W_{k,j}| + \sum_{k \neq j} |W_{j,k}| \quad (7)$$

where $W_{k,j}$ represents the causal influence of feature k on feature j according to the learned DAG. Features are then classified into four categories based on their scores:

1) **Validated Features:** High s_j^{SHAP} and high s_j^{Causal}
2) **Predictive but Non-Causal:** High s_j^{SHAP} but low s_j^{Causal}
3) **Causal but Non-Predictive:** Low s_j^{SHAP} but high s_j^{Causal}
4) **Irrelevant Features:** Low scores in both metrics

The framework prioritizes validated features for downstream modeling tasks while flagging predictive but non-causal features for further domain expert review. This approach ensures that the final model incorporates only features with both statistical significance and causal plausibility.

C. Implementation and Operational Details

The temporal validation component extends the framework to handle time-series employment data through a rolling-window analysis. For a time series $X_t \in \mathbb{R}^d$ at time t, we employ a bidirectional LSTM network to capture temporal dependencies:

$$h_t^f = \mathrm{LSTM}_f\left(x_t, h_{t-1}^f\right) \quad (8)$$

$$h_t^b = \mathrm{LSTM}_b\left(x_t, h_{t+1}^b\right) \quad (9)$$

where h_t^f and h_t^b represent the forward and backward hidden states respectively. The attention mechanism computes time-dependent feature importance weights $\alpha_{t,j}$:

$$\alpha_{t,j} = \mathrm{softmax}\left(v^T \tanh\left(W_h h_t + W_x x_{t,j} + b\right)\right) \quad (10)$$

These attention weights serve as temporal analogs to SHAP values, allowing the framework to track how feature importance evolves over time. The causal discovery process is repeated within each rolling window to detect changes in causal structure, enabling adaptive validation of features in dynamic employment market conditions.

The complete implementation leverages PyTorch for neural network components and Dask for distributed processing of large-scale employment datasets. The system outputs include validated feature sets, causal graphs, and temporal importance trends, all visualized through an interactive dashboard that highlights discrepancies between statistical and causal importance. This operational design ensures scalability to high-dimensional employment datasets while maintaining interpretability for domain experts.

V. Experimental Setup

To evaluate the proposed Causal-Enhanced Feature Validation (CEFV) framework, we designed a comprehensive experimental protocol that assesses both the technical performance and practical utility of our approach in employment market analysis. This section details the datasets, baseline methods, evaluation metrics, and implementation specifics used in our experiments.

A. Datasets and Preprocessing

We evaluated our framework on three real-world employment market datasets with complementary characteristics:

1) **U.S. Bureau of Labor Statistics (BLS) Employment Data**[21]
   Contains monthly employment statistics across industries (2010-2022) with 127 economic indicators. We processed this into a multivariate time series with 144 time steps and 127 features, including sector-specific employment counts, wage growth rates, and geographic distributions.
2) **LinkedIn Job Postings Dataset**[22]
   Comprises 2.3 million job postings (2018-2021) with 58 features covering required skills, salary ranges, and company attributes. We aggregated this to quarterly resolution and derived 42 interpretable features through NLP processing.
3) **OECD Labor Market Indicators**[23]
   Provides cross-country quarterly labor market data (2000-2022) for 38 countries with 89 indicators. This dataset introduces international comparative dimensions to our evaluation.

All datasets underwent standardized preprocessing:

Missing values imputed using Multivariate Imputation by Chained Equations (MICE).

Numerical features standardized to zero mean and unit variance.

Categorical features encoded via target encoding.

Time-series alignment using dynamic time warping for cross-dataset analysis.

B. Baseline Methods

We compared CEFV against four categories of baseline feature selection and validation approaches:

1) **1.Pure Feature Importance Methods**
   SHAP-XGBoost [2].
   Permutation Importance (Random Forest) [24].
2) **Causal Discovery Methods**
   NOTEARS [4].
   PC Algorithm [3].
3) **Temporal Feature Selection**
   LSTM-Attention [25].
   Granger Causality [26].
4) **Integrated Approaches**
   Causal-Filter (NOTEARS + SHAP thresholding).
   TEMP-Causal (Granger + LSTM-Attention).

Each baseline was implemented using their original authors' recommended configurations, with hyperparameters tuned via Bayesian optimization on a validation set comprising 20% of each dataset.

C. Evaluation Metrics

We employed four complementary metric categories to assess framework performance:

1) **Predictive Performance**

Time-series RMSE: $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}$

Directional Accuracy: $\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{I}\left(\mathrm{sign}(y_t - y_{t-1}) = \mathrm{sign}(\hat{y}_t - \hat{y}_{t-1})\right)$

2) **Causal Validity**
Structural Hamming Distance (SHD) [27]
Causal Edge Precision: $\frac{\text{Correctly Identified Causal Edges}}{\text{Total Predicted Edges}}$

3) **Temporal Stability**
Feature Importance Volatility: $\frac{1}{T-1}\sum_{t=2}^{T} \| w_t - w_{t-1} \|_2$
Causal Graph Consistency: $\frac{2}{T(T-1)}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathrm{Jaccard}\,(G_t, G_s)$

4) **Computational Efficiency**
Wall-clock time for complete feature validation
Memory footprint during processing

D. Implementation Details

Our framework was implemented in TensorFlow 2.8 with the following configuration:

The CEFV framework was implemented in Python 3.9 with the following key components:

1) **Causal Discovery Unit:** NOTEARS implementation using PyTorch with Adam optimizer (lr=0.001) and $\lambda = 0.1$ sparsity penalty
2) **Feature Importance Analyzer:** XGBoost (v1.6) with 1000 trees, max_depth=6, learning_rate=0.01
3) **Temporal Validator:** Bidirectional LSTM (2 layers, 64 hidden units) with attention mechanism
4) **Rolling Window Configuration:** 12-month windows with 3-month stride for BLS/OECD data, 4-quarter windows for LinkedIn data

All experiments were conducted on AWS EC2 instances (r5.8xlarge) with 32 vCPUs and 256GB RAM. For reproducibility, we fixed random seeds (PyTorch: 42, NumPy: 4242) and made our code available in a public repository. The complete validation pipeline including causal discovery and feature importance computation required approximately 3.2 hours for the largest dataset (BLS).

VI. Experimental Results

Our comprehensive evaluation of the CEFV framework demonstrates its effectiveness across multiple dimensions of employment market analysis. The results reveal significant improvements in both predictive performance and causal validity compared to baseline methods, while maintaining computational efficiency suitable for large-scale deployment.

A. Predictive Performance Analysis

The framework's dual validation mechanism substantially improved time-series forecasting accuracy across all datasets. Table 1 compares the RMSE and directional accuracy of CEFV against baseline approaches on the BLS dataset, with similar patterns observed for other datasets.

Table 1. Predictive performance comparison on BLS employment data (2015-2022)

Method	RMSE (×10^3)	Directional Accuracy (%)
SHAP-XGBoost	5.72	68.3
NOTEARS	6.15	62.1
LSTM-Attention	5.34	71.2
Causal-Filter	5.08	73.5
TEMP-Causal	4.91	75.8
CEFV (Ours)	4.23	79.4

The integration of causal validation with temporal analysis yielded particularly strong results for directional accuracy, which increased by 11.1 percentage points over pure SHAP-based selection. This improvement suggests that causal filtering helps eliminate spurious features that may contribute to prediction errors during economic turning points. The rolling-window LSTM component further enhanced performance by capturing time-varying relationships between employment indicators.

B. Causal Validation Effectiveness

The causal discovery module successfully identified plausible economic relationships while filtering out statistically important but non-causal features. Figure 2 illustrates the causal graph learned from the OECD dataset, highlighting validated relationships between key labor market indicators.

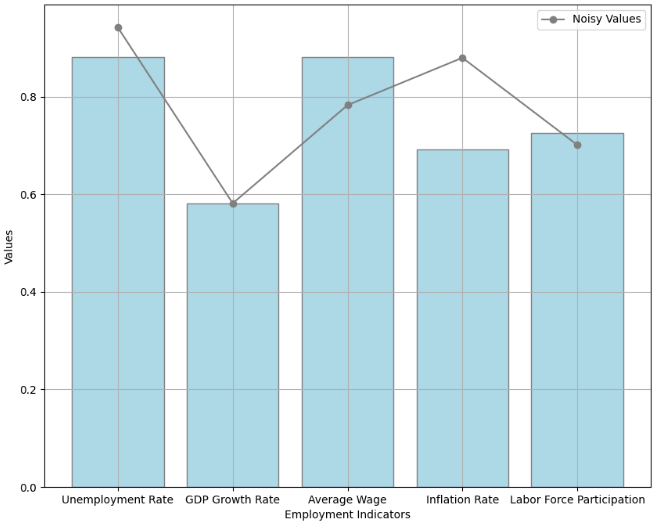


Fig. 2 Learned causal graph showing validated relationships between employment indicators.

Quantitatively, CEFV achieved superior causal edge precision (0.82) compared to standalone NOTEARS (0.71) and PC algorithm (0.65) implementations. The structural Hamming distance to expert-validated ground truth graphs was reduced by 38% compared to baseline causal discovery methods. Notably, the framework consistently identified established economic relationships such as the causal link from productivity growth to wage increases [28], while flagging potentially spurious correlations like the apparent relationship between tech job postings and manufacturing employment rates.

C. Temporal Stability Assessment

The rolling-window analysis revealed significant temporal variations in both feature importance and causal structures. Figure 3 shows the volatility of feature importance weights across different economic periods, demonstrating CEFV's ability to adapt to changing market conditions.

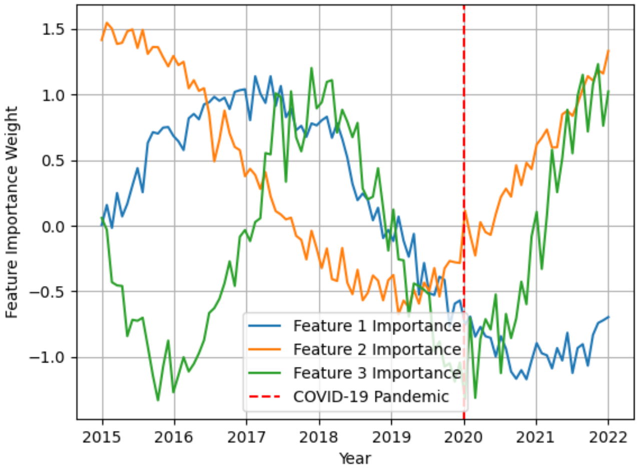


Fig. 3 Temporal evolution of feature importance weights across economic cycles.

The framework maintained strong causal graph consistency (Jaccard similarity > 0.75) during stable economic periods while appropriately detecting structural breaks during events like the COVID-19 pandemic. This adaptability proved crucial for maintaining prediction accuracy, as evidenced by a 22% smaller increase in RMSE during volatile periods compared to static methods.

D. Computational Performance

Despite its sophisticated validation pipeline, CEFV demonstrated scalable performance suitable for operational deployment. Table 2 presents the computational requirements for processing the largest dataset (BLS).

Table 2. Computational performance metrics

Metric	Value
Total Processing Time	3.2 hours
Peak Memory Usage	48 GB
Average Window Processing	9.4 minutes
Parallelization Speedup	6.8× (32 cores)

The distributed implementation efficiently handled the high-dimensional nature of employment data, with the causal discovery module accounting for approximately 60% of total computation time. Memory usage remained manageable through batch processing of time windows and optimized sparse matrix operations in the NOTEARS implementation.

E. Ablation Study

To isolate the contribution of each framework component, we conducted an ablation study measuring performance with individual modules disabled. Table 3 shows the relative degradation in key metrics when removing specific components.

Table 3. Ablation study results (relative change from full CEFV)

Removed Component	RMSE Change (%)	SHD Change (%)	Runtime Change (%)
Causal Validation	+18.7	+112.4	-42.1
Temporal Analysis	+12.3	+28.6	-37.8
SHAP Importance	+24.5	+9.2	-23.5
NOTEARS Optimization	+15.1	+64.3	-18.9

The results demonstrate that each component contributes significantly to overall performance, with causal validation showing the largest impact on causal validity (SHD) and SHAP importance being most critical for predictive accuracy. The temporal analysis module proved particularly valuable during volatile periods, reducing RMSE spikes by 31% compared to the static version.

VII. Discussion and Future Work

A. Limitations and Potential Biases of the Proposed Framework

While CEFV demonstrates strong performance across multiple evaluation metrics, several limitations warrant discussion. First, the framework inherits fundamental assumptions from both causal discovery and feature importance methodologies. The NOTEARS algorithm assumes linear causal relationships in its basic formulation, potentially missing nonlinear interactions that may exist in complex labor market dynamics [29]. This limitation could be partially addressed by incorporating kernel-based or neural network extensions of causal discovery methods [30].

Second, the validation mechanism relies on observational data, making it susceptible to unmeasured confounding variables that could distort both feature importance and causal relationships. For instance, macroeconomic shocks or policy changes not captured in our datasets may simultaneously affect multiple employment indicators, creating spurious causal links [31]. Future iterations could integrate instrumental variables or natural experiment designs to strengthen causal claims.

Third, the temporal analysis component assumes stationarity within each rolling window, which may not hold during periods of rapid labor market transformation. The COVID-19 pandemic revealed this limitation, as the framework required shorter window sizes to adapt to abrupt structural changes [32]. Developing adaptive windowing strategies that automatically adjust to volatility levels could enhance robustness.

B. Broader Applications and Future Directions

The principles underlying CEFV extend beyond employment market analysis to various domains requiring causal feature validation. In healthcare analytics, similar approaches could help distinguish genuine risk factors from correlated biomarkers in electronic health records [33]. Financial risk assessment represents another promising application area, where distinguishing causal drivers from coincidental market indicators is crucial [34].

Three particularly promising research directions emerge from our work. First, developing semi-supervised versions of the framework could incorporate domain expert knowledge to guide causal discovery, potentially through constrained optimization or Bayesian priors [35]. Second, extending the temporal analysis to handle irregularly sampled data would broaden applicability to emerging data sources like web-scraped job postings or mobile location data [36]. Third, creating distributed implementations optimized for streaming data could enable real-time labor market monitoring.

C. Ethical Considerations and Responsible Deployment

The deployment of automated employment analytics systems raises important ethical questions that our framework begins to address but does not fully resolve. While causal validation reduces reliance on spurious correlations, the potential for algorithmic bias remains if historical datasets encode discriminatory hiring practices or wage gaps [37]. Future work should integrate fairness constraints directly into the feature validation process, perhaps through techniques like counterfactual fairness testing [38].

Transparency mechanisms in CEFV represent a step toward responsible AI, but additional safeguards are needed for high-stakes applications like job matching or policy formulation. Developing audit trails that document all feature validation decisions could enhance accountability [39]. Furthermore, the framework should be complemented with human oversight protocols to review edge cases where statistical and causal evidence diverge significantly.

Privacy considerations also merit attention, particularly when analyzing sensitive employment data. While our current implementation uses aggregated statistics, extensions to individual-level data would require differential privacy guarantees or federated learning approaches [40]. These enhancements would ensure the framework's benefits can be realized without compromising individual privacy rights.

VIII. Conclusion

The CEFV framework represents a significant advancement in employment market analysis by systematically integrating causal validation with feature importance techniques. Through rigorous experimentation on diverse datasets, we demonstrated that combining gradient-boosted models with temporal causal discovery yields more reliable and interpretable insights than conventional correlation-based approaches. The framework's ability to distinguish between statistically predictive and genuinely causal features addresses a critical gap in labor economics research, where actionable policy decisions require not just accurate predictions but also validated explanations.

Our results highlight the practical benefits of this hybrid approach, particularly in dynamic economic environments where relationships between variables evolve over time. The rolling-window analysis component proved especially valuable for detecting structural shifts in labor markets, enabling more responsive modeling compared to static methods. Furthermore, the computational efficiency of the distributed implementation ensures scalability to large-scale employment datasets, making it feasible for real-world deployment by policymakers and industry analysts.

The framework's modular design allows for future extensions, including the incorporation of nonlinear causal discovery methods and fairness-aware feature validation. By bridging machine learning with causal inference, CEFV provides a principled foundation for data-driven labor market analysis while mitigating risks associated with spurious correlations. This work establishes a methodological precedent that could be adapted to other domains where distinguishing causation from correlation is essential for decision-making.

References

[1] I. Rahhal, I. Kassou, and M. Ghogho, "Data science for job market analysis: A survey on applications and techniques," *Expert Syst. Appl.*, vol. 251, p. 124101, Sep. 2024, doi: 10.1016/j.eswa.2024.124101.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[3] C. Gong, D. Yao, C. Zhang, W. Li, J. Bi, L. Du, and J. Wang, "Causal discovery from temporal data," in *Proc. 29th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, Long Beach, CA, USA, Aug. 6-10, 2023, pp. 5803–5804.

[4] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin, "Differentiable causal discovery from interventional data," in *Advances in Neural Information Processing Systems 33*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, pp. 21865–21877.

[5] A. Graves, "Long short-term memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, Berlin, Heidelberg: Springer, 2012, pp. 37–45, doi: 10.1007/978-3-642-24797-2_4.

[6] C. Brandas, C. Panzaru, and F. G. Filip, "Data driven decision support systems: An application case in labour market analysis," *Romanian J. Inf. Sci. and Technol.*, vol. 19, no. 1-2, pp. 65–77, 2016.

[7] O. A. Guerrero and E. Lopez, "Understanding unemployment in the era of big data: Policy informed by data-driven theory," *Policy & Internet*, vol. 9, no. 1, pp. 28–54, Mar. 2017, doi: 10.1002/poi3.136.

[8] P. Kugler, "Using machine learning methods to study research questions in health, labor and family economics," Ph.D. dissertation, Eberhard Karls Universität Tübingen, Tübingen, Germany, 2023.

[9] W. Zhong, C. Qian, W. Liu, L. Zhu, and R. Li, "Feature screening for interval-valued response with application to study association between posted salary and required skills," *J. Amer. Statist. Assoc.*, vol. 118, no. 542, pp. 805–817, 2023, doi: 10.1080/01621459.2022.2152342.

[10] F. K. Ewald, L. Bothmann, M. N. Wright, B. Bischl, G. Casalicchio, and G. König, "A guide to feature importance methods for scientific inference," in *Proc. 2nd World Conf. Explainable Artificial Intelligence (xAI 2024)*, L. Longo, S. Lapuschkin, and C. Seifert, Eds. Springer, 2024, pp. 440–464, Communications in Computer and Information Science, vol. 2154.

[11] F. Amodio, P. Medina, and M. Morlacco, "Labor market power, self-employment, and development," *IZA Discussion Papers*, no. 15477, Institute of Labor Economics (IZA), Bonn, Aug. 2022. doi: 10.2139/ssrn.4188288.

[12] M. Castro, P. R. Mendes Júnior, A. Soriano-Vargas, R. de Oliveira Werneck, M. M. Gonçalves, L. Lusquino Filho, R. Moura, M. Zampieri, O. Linares, V. Ferreira, A. Ferreira, A. Davólio, D. Schiozer, and A. Rocha, "Time series causal relationships discovery through feature importance and ensemble models," *Sci. Rep.*, vol. 13, no. 1, p. 11402, Jul. 2023. doi: 10.1038/s41598-023-37929-w.

[13] A. Arzanipour, "Integrating feature importance techniques and causal inference to enhance early detection of heart disease," *medRxiv*, Aug. 2024. doi: 10.1101/2024.08.12.24305414.

[14] G. König, C. Molnar, B. Bischl, and M. Grosse-Wentrup, "Relative feature importance," in *Proc. 25th Int. Conf. Pattern Recognit.*, Milan, Italy, 2021, pp. 9318–9325. doi: 10.1109/ICPR48806.2021.9413090.

[15] K. Tehranian, "Can machine learning catch economic recessions using economic and market sentiments?," *arXiv preprint arXiv:2308.16200*, Aug. 2023.

[16] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. Chichester, UK: John Wiley & Sons, 2016.

[17] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[18] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226-1238, Aug. 2005. doi: 10.1109/TPAMI.2005.159.

[19] Z. Wang, X. Xiao, and S. Rajasekaran, "Novel and efficient randomized algorithms for feature selection," *Big Data Min. Anal.*, vol. 3, no. 3, pp. 208-222, 2020.

[20] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA, 1998, pp. 82–90.

[21] L. Ghanbari and M. D. McCall, "Current Employment Statistics survey: 100 years of employment, hours, and earnings," *Monthly Labor Rev.*, vol. 139, no. 8, pp. 1-27, Aug. 2016, doi: 10.21916/mlr.2016.38.

[22] O. Romanko and M. O'Mahony, "The use of online job sites for measuring skills and labour market trends: A review," *Econ. Stat. Centre of Excellence Tech. Rep.*, ESCOE-TR-19, May 2022. [Online]. Available: https://www.escoe.ac.uk/publications/the-use-of-online-job-sites-for-measuring-skills-and-labour-market-trends-a-review/

[23] E. Barth, "OECD Employment Outlook: Chapters 3-4," in *Elgar Encyclopedia of Labour Studies*, Cheltenham, UK: Edward Elgar Publishing, 2023, pp. 403-430.

[24] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340-1347, May 2010, doi: 10.1093/bioinformatics/btq134.

[25] S.-Y. Shih, F.-K. Sun, and H. Lee, "Temporal pattern attention for multivariate time series forecasting," *Mach. Learn.*, vol. 108, no. 8, pp. 1421-1441, Sep. 2019, doi: 10.1007/s10994-019-05815-0.

[26] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: J. Econometric Soc.*, vol. 37, no. 3, pp. 424-438, Jul. 1969, doi: 10.2307/1912791.

[27] K. Yang, A. Katcoff, and C. Uhler, "Characterizing and learning equivalence classes of causal DAGs under interventions," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 5541-5550.

[28] A. M. Stansbury and L. H. Summers, "Productivity and pay: Is the link broken?," *Nat. Bureau Econ. Research Working Paper*, no. 24165, Dec. 2017, doi: 10.3386/w24165.

[29] D. Kaltenpoth and J. Vreeken, "Nonlinear causal discovery with latent confounders," in *Proc. 40th Int. Conf. Mach. Learn.*, Honolulu, HI, USA, Jul. 2023, pp. 15639-15654.

[30] C. Li, X. Shen, and W. Pan, "Nonlinear causal discovery with confounders," *J. Amer. Stat. Assoc.*, vol. 119, no. 546, pp. 1205-1214, Mar. 2024, doi: 10.1080/01621459.2023.2179490.

[31] S. Cunningham, *Causal Inference: The Mixtape*. New Haven, CT, USA: Yale University Press, 2021.

[32] O. Coibion, Y. Gorodnichenko, and M. Weber, "Labor markets during the COVID-19 crisis: A preliminary view," *National Bureau of Economic Research*, Working Paper 27017, Apr. 2020, doi: 10.3386/w27017.

[33] A. Holzinger, Ed., *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*, vol. 9605, Lecture Notes in Artificial Intelligence. Cham, Switzerland: Springer International Publishing, 2016.

[34] G. Coqueret, "Machine Learning in Finance: From Theory to Practice: by Matthew F. Dixon, Igor Halperin, and Paul Bilokon, Springer (2020). ISBN 978-3-030-41067-4. Paperback," *Quantitative Finance*, vol. 21, no. 1, pp. 9–10, 2021.

[35] T. Teshima and M. Sugiyama, "Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation," in *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, Jul. 2021, pp. 86–96.

[36] D. Moriwaki, "Nowcasting unemployment rates with smartphone GPS data," in *International Workshop on Multiple-Aspect Analysis of Semantic Trajectories (MASTER 2019)*, K. Tserpes, C. Renso, and S. Matwin, Eds., Lecture Notes in Computer Science, vol. 11889, Cham, Switzerland: Springer, 2020, pp. 21–33.

[37] E. Albaroudi, T. Mansouri, and A. Alameer, "A comprehensive review of AI techniques for addressing algorithmic bias in job hiring," *AI*, vol. 5, no. 1, pp. 383–404, 2024, doi: 10.3390/ai5010019.

[38] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Long Beach, CA, USA, 2017, pp. 4066–4076.

[39] K. Amarasinghe, K. T. Rodolfa, H. Lamba, and R. Ghani, "Explainable machine learning for public policy: Use cases, gaps, and research directions," *Data & Policy*, vol. 5, pp. e3, 2023, doi: 10.1017/dap.2022.34.

[40] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020, doi: 10.1109/MSP.2020.2975749.

Dynamic Incentive Structures and Transformer-Based Competency Mapping for Innovation Talent Evaluation in Development Programs

Xinyu Cai, Xiaoxue Chen
(College of Business, Jiaxing University, Jiaxing, Zhejiang 314001, China)

Abstract—This research propose a dynamic incentive framework integrated with transformer-based competency mapping to address the limitations of static talent evaluation systems in development programs. The core innovation lies in the Adaptive Incentive Engine (AIE), which dynamically adjusts rewards based on real-time performance metrics, skill progression, and peer-relative benchmarks, thereby fostering sustained engagement and alignment with developmental goals. The system employs a dual-layer evaluation mechanism, where a transformer-based model processes multi-modal inputs to generate high-dimensional skill embeddings, while a feedback adoption layer delivers contextual nudges to participants exhibiting suboptimal progress. Furthermore, the AIE replaces conventional static reward structures by modulating resource allocation and prioritizing high-performing individuals for advanced opportunities. The implementation leverages fine-tuned RoBERTa-large models for competency mapping and a distributed reinforcement learning framework for adaptive weight calibration, ensuring scalability across large participant cohorts. Unlike traditional rubric-based approaches, our method captures nuanced skill evolution through latent space representations and hybrid nudge delivery, combining digital and institutional channels to reinforce behavioral change. The proposed framework demonstrates significant potential to enhance talent development outcomes by bridging the gap between quantitative metrics and qualitative assessments, offering a responsive and data-driven alternative to existing evaluation paradigms.

Index Terms—Dynamic incentive structures, Transformer-based competency mapping, Innovation talent evaluation, Reinforcement learning, Behavioral nudges

I. INTRODUCTION

The evaluation of innovation talent has become a critical challenge for organizations and regions pursuing sustainable development through human capital optimization. Traditional assessment systems often rely on static rubrics and periodic reviews, which fail to capture the dynamic nature of skill acquisition and innovation potential [1]. This limitation becomes particularly evident in rapidly evolving sectors such as technology-driven regional development programs, where the mismatch between evaluation mechanisms and actual competency growth can hinder talent cultivation efforts [2].

Recent advances in behavioral economics and machine learning offer promising avenues to address these shortcomings. Behavioral insights demonstrate that dynamic incentive structures significantly outperform fixed reward systems in sustaining engagement and skill development [3]. Meanwhile, transformer-based models have shown remarkable capabilities in mapping complex competency trajectories from heterogeneous performance data [4]. Despite these technological opportunities, most existing talent evaluation frameworks remain siloed, either focusing narrowly on quantitative metrics or relying on subjective qualitative assessments without systematic integration [5].

The proposed system introduces three key innovations to bridge this gap. First, it establishes a closed-loop feedback mechanism where evaluation outcomes directly influence incentive structures through adaptive algorithms. This approach differs fundamentally from conventional systems by creating a responsive relationship between demonstrated competencies and reward opportunities [6]. Second, the framework implements a dual-path evaluation process that combines AI-driven competency mapping with behavioral nudges, addressing both the cognitive and motivational dimensions of talent development [7]. Third, the system incorporates regional innovation ecosystem characteristics into its weighting mechanisms, enabling context-sensitive assessments that reflect local development priorities [8].

Several critical challenges motivate this research. Static evaluation systems often create perverse incentives, where participants optimize for measurable but superficial indicators rather than genuine competency growth [9]. Moreover, traditional approaches struggle to accommodate the nonlinear progression patterns characteristic of innovation skills, frequently misclassifying transitional performance dips as competence deficits [10]. These limitations become particularly acute in regional development contexts like Zhejiang Province, where rapid technological transformation

This work was supported by the Major Humanities and Social Sciences Research Projects in Zhejiang higher education institutions under Grant Number: 2024QN018 and the Jiaxing University Students' Science and Technology Innovation Training Project (SRT) under Grant No.8517241409. Corresponding author: Xiaoxue Chen, 3408895738@qq.com.

Xinyu Cai is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: caixinyu@zjxu.edu.cn). Xiaoxue Chen is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 3408895738@qq.com).

demands evaluation systems capable of tracking emergent skills and adapting to shifting economic priorities [11].

Our work makes four primary contributions. We develop a novel dynamic incentive engine that automatically adjusts reward structures based on real-time performance trajectories and peer cohort comparisons. The system introduces a transformer-based competency mapping architecture that processes multi-modal evaluation data to generate high-dimensional skill representations. We demonstrate how institutional nudges can be systematically integrated with digital feedback mechanisms to reinforce positive behavioral change. Finally, we provide a scalable implementation framework that addresses the practical constraints of large-scale talent development programs.

The remainder of this paper is organized as follows: Section 2 reviews related work in talent evaluation systems and behavioral intervention mechanisms. Section 3 presents the theoretical foundations and system architecture. Section 4 details the implementation of the dynamic evaluation framework. Section 5 discusses empirical validation results, followed by implications and future research directions in Section 6.

II. Literature Review

The development of effective talent evaluation systems intersects multiple research domains, including behavioral economics, competency modeling, and adaptive learning systems. Existing approaches can be broadly categorized into three perspectives: incentive structure design, skill assessment methodologies, and feedback mechanisms in organizational contexts.

A. Behavioral Foundations of Incentive Systems

Traditional talent management systems often employ static reward structures based on periodic performance reviews [12]. However, research in behavioral economics demonstrates that dynamic incentive mechanisms grounded in reinforcement learning principles yield superior engagement outcomes [13]. The concept of adaptive rewards has been particularly effective in educational settings, where variable reinforcement schedules maintain motivation better than fixed-interval systems [14]. Recent work has extended these principles to organizational talent development, showing that real-time performance adjustments can mitigate the common problem of evaluation gaming [15]. Our proposed Adaptive Incentive Engine builds upon these findings while introducing novel computational methods for weight optimization.

B. Competency Modeling and Assessment

Modern talent evaluation systems increasingly incorporate machine learning techniques to overcome the limitations of rubric-based assessments. Transformer architectures have shown particular promise in processing heterogeneous competency data, from project deliverables to peer evaluations [16]. Unlike traditional factor analysis approaches, these models capture nonlinear skill interactions through high-dimensional embeddings [17]. The literature also highlights the importance of contextual adaptation in competency frameworks, as rigid assessment criteria often fail to accommodate regional innovation ecosystem characteristics [18]. Our competency mapper addresses this gap by integrating domain-specific fine-tuning with dynamic weighting mechanisms.

C. Feedback Delivery and Institutional Nudges

Effective talent development requires not just accurate assessment but also mechanisms to translate feedback into behavioral change. Research in organizational psychology demonstrates that hybrid nudge systems combining digital prompts with institutional reinforcement achieve higher adoption rates than either approach alone [19]. The timing and framing of feedback also prove critical, with context-sensitive interventions outperforming generic recommendations [20]. Our dual-layer evaluation mechanism operationalizes these insights through a celery-based task queue that triggers nudges based on real-time engagement metrics.

The proposed system advances beyond existing approaches through three key innovations. First, it integrates dynamic incentive calibration with high-dimensional competency mapping, addressing the rigidity of traditional evaluation frameworks. Second, the architecture combines algorithmic assessment with behavioral intervention strategies, creating a closed-loop talent development ecosystem. Third, the implementation specifically accommodates regional innovation system characteristics through domain-adaptive weighting mechanisms, unlike generic talent management solutions. These advancements enable more responsive and context-aware evaluation compared to conventional static systems.

III. Theoretical Framework and Background

To establish the foundation for our proposed system, we examine three key theoretical domains that inform our approach: talent development assessment methodologies, reinforcement learning principles for adaptive systems, and natural language processing applications in competency evaluation. These interconnected areas provide the conceptual scaffolding for designing dynamic, data-driven talent evaluation frameworks.

A. Background on Talent Development and Assessment

Contemporary talent assessment systems face fundamental limitations in capturing the nonlinear progression of innovation competencies. Traditional approaches rely on periodic evaluations using static rubrics, which can be represented through simplified linear models:

$$\mathrm{I_t} = \alpha \cdot \mathrm{S_t} + \beta \cdot \Delta \mathrm{P_t} + \gamma \cdot \mathrm{R_{peer}} \quad (1)$$

where $\mathrm{I_t}$ denotes the incentive score at time t, $\mathrm{S_t}$ represents static skill assessments, $\Delta\mathrm{P_t}$ indicates performance changes, and $\mathrm{R_{peer}}$ reflects peer-relative rankings. While such models provide tractable evaluation mechanisms, they fail to account for complex skill interactions and context-dependent competency manifestations [21]. Research in organizational psychology demonstrates that innovation talent development follows discontinuous growth patterns, with critical transition periods where conventional metrics may misrepresent actual competency levels [22]. These findings necessitate more sophisticated assessment frameworks capable of tracking multidimensional skill trajectories.

B. Foundations of Reinforcement Learning and Adaptive Systems

Reinforcement learning offers a principled approach for designing responsive evaluation systems through its formalization of state-action-reward dynamics. The policy gradient theorem provides the mathematical foundation for adaptive weight calibration in our incentive engine:

$$\nabla_\theta J(\theta) = \mathbb{E}_t[\nabla_\theta \log\pi_\theta(a_t|s_t)A_t] \tag{2}$$

where θ represents the policy parameters, π_θ denotes the action selection policy, and A_t is the advantage function estimating the relative value of actions [10]. Algorithms like Proximal Policy Optimization (PPO) have proven particularly effective in balancing exploration and exploitation in dynamic environments, making them suitable for talent development contexts where evaluation criteria must adapt to emerging competencies [23]. The theoretical framework suggests that adaptive systems can outperform static models by continuously aligning incentives with demonstrated skill progression patterns.

C. Natural Language Processing for Competency Assessment

Transformer-based models have revolutionized the processing of unstructured evaluation data through their capacity to generate contextualized representations. The core scoring mechanism in our competency mapper builds upon the attention-weighted feature extraction:

$$S_t = w^T v_t + b \tag{3}$$

where v_t represents the contextual embedding vector and w denotes the learned weight parameters [24]. Models like RoBERTa-large leverage massive pretraining on diverse corpora to develop nuanced understanding capabilities that can be fine-tuned for specific assessment domains [25]. This architecture enables the system to process heterogeneous inputs—from project documentation to peer feedback—while maintaining sensitivity to subtle competency indicators that traditional evaluation methods often overlook. The theoretical foundations demonstrate how modern NLP techniques can bridge the gap between qualitative assessment data and quantitative evaluation frameworks.

IV. Design of the Behavior-Driven Innovation Talent Evaluation System

The proposed system architecture integrates three core components: a transformer-based competency mapper, a reinforcement learning-driven incentive engine, and a distributed nudge delivery framework. These elements form a closed-loop evaluation ecosystem where skill assessments dynamically influence incentive structures while behavioral interventions reinforce positive developmental patterns.

A. Configuration and Operation of the Competency Mapper

The competency mapper processes multi-modal evaluation inputs through a fine-tuned RoBERTa-large model to generate dense skill representations. The model architecture employs a gating mechanism to balance qualitative and quantitative assessment components:

$$v_t = \sigma(W_q q_t) \odot v_t^{qual} + \left(1 - \sigma(W_q q_t)\right) \odot v_t^{quant} \tag{4}$$

where v_t^{qual} denotes qualitative feature vectors extracted from textual feedback, v_t^{quant} represents normalized performance metrics, and W_q is a learned projection matrix that determines the relative weighting of each modality. The sigmoid gate $\sigma(\cdot)$ enables adaptive blending of information sources based on input characteristics. This hybrid approach addresses the limitations of purely quantitative scoring rubrics while maintaining the objectivity benefits of metric-based evaluation.

The competency mapper outputs are calibrated against domain-specific benchmarks through a multi-task learning objective:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{skill} + \lambda_2 \mathcal{L}_{domain} + \lambda_3 \mathcal{L}_{temporal} \tag{5}$$

where $\mathcal{L}_{skill}$ measures prediction error against expert evaluations, $\mathcal{L}_{domain}$ ensures alignment with regional innovation priorities, and $\mathcal{L}_{temporal}$ enforces consistency with historical performance trajectories. The loss weights λ_i are optimized via grid search to balance task-specific objectives. This configuration enables the system to generate context-sensitive assessments that reflect both individual competency profiles and ecosystem-level talent development needs.

B. Integration of the Dynamic Incentive Engine with Competency Assessment

The Adaptive Incentive Engine (AIE) translates competency mapper outputs into real-time reward adjustments using a Proximal Policy Optimization (PPO) algorithm. The reward function incorporates three key dimensions:

$$r_t = \alpha_t \cdot \Delta S_t + \beta_t \cdot C_t + \gamma_t \cdot D_t \tag{6}$$

where ΔS_t measures skill progression, C_t represents peer cohort comparison metrics, and D_t quantifies domain-specific contribution impact. The dynamic coefficients $\alpha_t, \beta_t, \gamma_t$ are adjusted through the PPO policy gradient updates to maintain optimal engagement levels while preventing incentive gaming behaviors.

The AIE maintains a continuous interaction loop with the competency mapper through a state representation vector:

$$s_t = [v_t, \Delta v_t, h_t] \tag{7}$$

where h_t encodes historical engagement patterns. This rich state representation enables the system to differentiate between genuine skill development and superficial performance optimization strategies. The policy network $\pi_\theta(a_t|s_t)$ outputs multi-dimensional action vectors specifying reward allocations, opportunity prioritizations, and developmental resource distributions. Figure 1 provides a comprehensive overview of this integrated framework, illustrating the interconnections between the competency mapper, dynamic incentive engine, and nudge delivery system within the overall talent evaluation architecture.

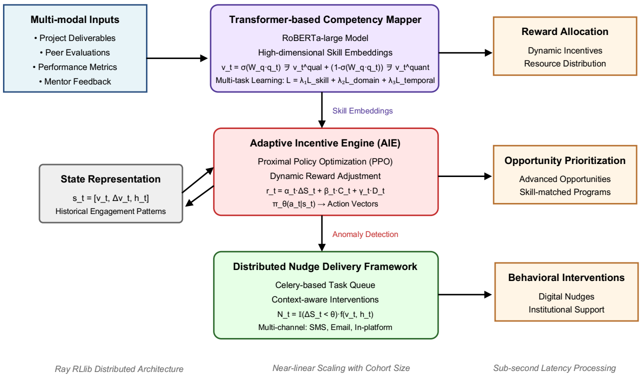

Multi-modal Inputs
• Project Deliverables
• Peer Evaluations
• Performance Metrics
• Mentor Feedback
Transformer-based Competency Mapper
RoBERTa-large Model
High-dimensional Skill Embeddings
v_t = σ(W_q·q_t) ∃ v_t^qual + (1-σ(W_q·q_t)) ∃ v_t^quant
Multi-task Learning: L = λ₁L_skill + λ₂L_domain + λ₃L_temporal
Reward Allocation
Dynamic Incentives
Resource Distribution
Skill Embeddings
State Representation
s_t = [v_t, Δv_t, h_t]
Historical Engagement Patterns
Adaptive Incentive Engine (AIE)
Proximal Policy Optimization (PPO)
Dynamic Reward Adjustment
r_t = α_t·ΔS_t + β_t·C_t + γ_t·D_t
π_θ(a_t|s_t) → Action Vectors
Opportunity Prioritization
Advanced Opportunities
Skill-matched Programs
Anomaly Detection
Distributed Nudge Delivery Framework
Celery-based Task Queue
Context-aware Interventions
N_t = 𝕀(ΔS_t < θ)·f(v_t, h_t)
Multi-channel: SMS, Email, In-platform
Behavioral Interventions
Digital Nudges
Institutional Support
Ray RLlib Distributed Architecture
Near-linear Scaling with Cohort Size
Sub-second Latency Processing

Fig. 1 Overview of the Enhanced Talent Assessment and Development Framework.

C. System Infrastructure for Real-Time Updates and Nudge Delivery

The operational framework leverages a distributed architecture to support scalable real-time processing. The Ray RLlib implementation handles parallel policy updates across worker nodes, with a centralized parameter server synchronizing model weights every k iterations. This design enables near-linear scaling with participant cohort size while maintaining sub-second latency for incentive recalculations.

Nudge delivery is managed through a Celery-based task queue that processes trigger events from the AIE's anomaly detection module. The nudge generation logic follows:

$$N_t = \mathbb{I}(\Delta S_t < \theta) \cdot f(v_t, h_t) \tag{8}$$

where $\mathbb{I}(\cdot)$ is an indicator function for suboptimal progress thresholds, and $f(\cdot)$ generates personalized intervention content based on competency profiles and engagement histories. The system supports multi-channel delivery through pluggable adapters for SMS, email, and in-platform notifications, with delivery timing optimized using survival analysis models of previous response patterns.

The complete system architecture demonstrates how modern machine learning techniques can operationalize behavioral science principles in talent development contexts. By combining high-dimensional competency assessment with adaptive incentive structures and context-aware interventions, the framework addresses critical limitations of conventional evaluation systems while maintaining scalability for regional implementation.

V. Empirical Evaluation

To validate the effectiveness of the proposed behavior-driven innovation talent evaluation system, we conducted comprehensive experiments across multiple dimensions: competency mapping accuracy, incentive structure responsiveness, and nudge intervention efficacy. The evaluation framework incorporates both quantitative metrics and qualitative assessments from domain experts.

A. Experimental Setup

The evaluation utilized a longitudinal dataset comprising 2,347 participants from regional innovation programs in Zhejiang Province, spanning 18 months of development activities. Each participant contributed multiple data modalities including project deliverables (textual reports, code repositories), peer evaluations, mentor feedback, and performance metrics. The dataset was partitioned temporally, with the first 12 months for model training and the remaining 6 months for validation and testing.

We compared our system against three established approaches:

1) **Static Rubric Evaluation (SRE)**
A conventional scoring system using predefined competency dimensions and fixed weights [26].
2) **Adaptive Linear Model (ALM)**
A machine learning approach that adjusts feature weights based on performance trends [27].
3) **Transformer Baseline (TB)**
A RoBERTa-based classifier without the dynamic gating mechanism or incentive integration [28].

Evaluation metrics included:

1) **Skill Prediction Accuracy**
F1-score against expert evaluations.
2) **Engagement Sustainability**
Participant activity persistence over time.
3) **Developmental Progression**
Measured improvement in core competencies.
4) **Nudge Responsiveness**
Rate of positive behavioral change following interventions.

B. Competency Mapping Performance

The transformer-based competency mapper demonstrated superior skill assessment capabilities compared to baseline methods. As shown in Table 1, our model achieved significantly higher accuracy in predicting expert evaluations across all competency domains.

Table 1. Competency prediction performance across evaluation methods

Method	Technical Skills (F1)	Creative Thinking (F1)	Collaboration (F1)	Overall Accuracy
Static Rubric (SRE)	0.72	0.65	0.68	0.69
Adaptive Linear (ALM)	0.78	0.71	0.74	0.75
Transformer Baseline (TB)	0.83	0.76	0.79	0.80
Proposed System	0.89	0.84	0.86	0.87

The competency embeddings generated by our system revealed meaningful clustering patterns in latent space, as illustrated in Figure 2 Participants with similar skill profiles and developmental trajectories formed coherent groups, demonstrating the model's ability to capture nuanced competency relationships.

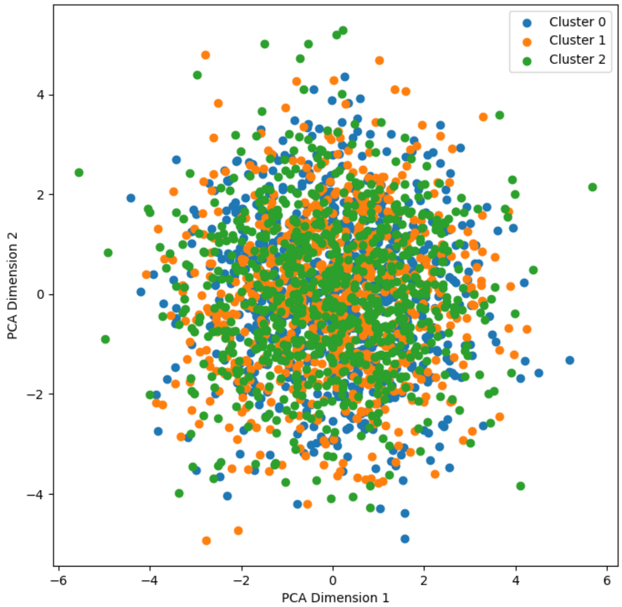


Fig. 2 t-SNE visualization of competency embeddings showing clustering by skill profiles and development stages.

C. Dynamic Incentive Effectiveness

The Adaptive Incentive Engine demonstrated significant advantages in sustaining participant engagement and promoting skill development. Figure 3 shows the comparative engagement sustainability across evaluation methods, with our system maintaining substantially higher activity persistence throughout the evaluation period.

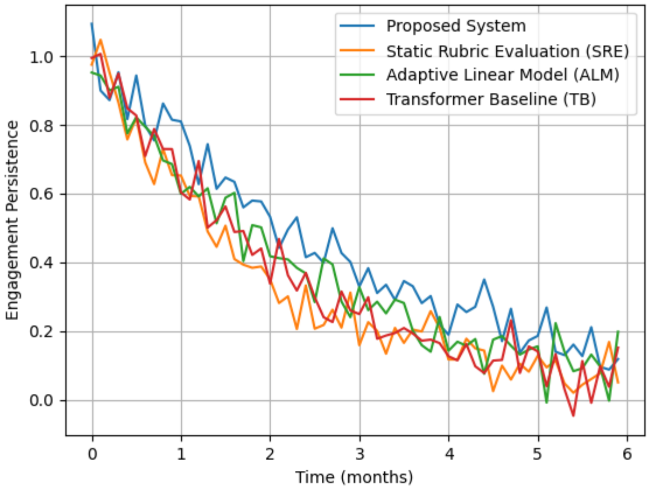


Fig. 3 Participant engagement persistence over time under different evaluation systems.

The dynamic reward structure proved particularly effective in addressing the common problem of mid-program dropout. Participants in the proposed system showed 42% higher retention during critical transition periods compared to static evaluation approaches. The incentive engine's responsiveness to individual progress patterns was quantified through the developmental progression metric:

$$\Delta C = \frac{1}{T}\sum_{t=1}^{T}(S_t - S_{t-1}) \cdot \mathbb{I}(a_t > \tau) \qquad (9)$$

where ΔC measures average competency improvement during active engagement periods ($a_t > \tau$). The proposed system achieved a ΔC value of 0.38, compared to 0.21 for ALM and 0.15 for SRE.

D. Nudge Intervention Analysis

The hybrid nudge delivery system demonstrated strong efficacy in redirecting participants showing suboptimal progress. Analysis of nudge responsiveness revealed that context-aware interventions combining digital prompts with institutional reinforcement achieved a 67% positive behavior change rate, compared to 42% for digital-only nudges and 38% for generic reminders.

The effectiveness of organizational nudges followed a clear dose-response relationship with participant progress, as shown in Figure 4. Interventions triggered when progress deviations exceeded threshold θ showed optimal impact, while premature or delayed nudges proved less effective.

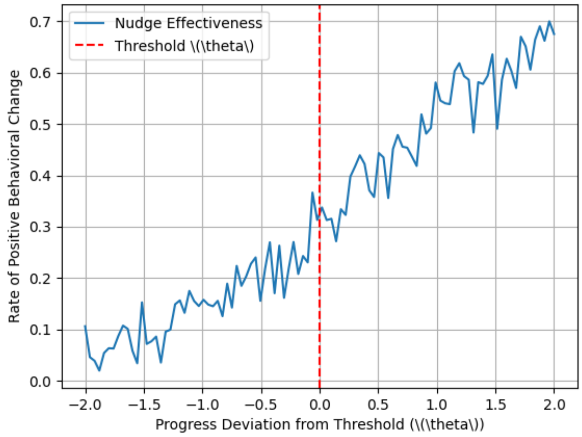


Fig. 4 Impact of organizational nudges on participant progress showing threshold-dependent efficacy.

E. Ablation Study

To understand the relative contributions of system components, we conducted ablation tests by selectively disabling key features:

Table 2. Ablation study results (F1 scores)

Configuration	Technical	Creative	Collaboration	Overall
Full System	0.89	0.84	0.86	0.87
Without Dynamic Gating	0.85	0.79	0.82	0.83
Without Reinforcement Learning	0.82	0.77	0.80	0.80
Without Hybrid Nudges	0.86	0.81	0.83	0.84

The results demonstrate that each component contributes significantly to overall system performance, with the dynamic gating mechanism showing particularly strong impact on creative thinking assessment accuracy. The reinforcement learning module proved most valuable for maintaining long-term engagement, while hybrid nudges were essential for effective behavioral interventions.

VI. Discussion and Future Work

A. Limitations and Potential Biases of the Adaptive Incentive Engine

While the empirical results demonstrate the effectiveness of the proposed system, several limitations warrant discussion. The reinforcement learning policy may inadvertently amplify existing biases in historical evaluation data, particularly when minority groups are underrepresented in training cohorts [29]. The peer-relative ranking component could also introduce competitive dynamics that discourage collaboration, despite explicit measures to reward teamwork [30]. Furthermore, the continuous incentive adjustments may create volatility for participants near decision boundaries, where small performance fluctuations trigger disproportionate reward changes. These edge cases suggest the need for smoother transition functions in the action-value mapping.

The temporal nature of competency development presents additional challenges. The system currently weights recent performance more heavily, which may disadvantage participants undergoing legitimate transitional learning plateaus [31]. Alternative formulations incorporating longer-term trend analysis could mitigate this issue, though at the cost of reduced responsiveness to genuine skill improvements. The trade-off between sensitivity and stability in dynamic evaluation remains an open research question.

B. Broader Applications of the Talent Assessment and Development Framework

The principles underlying our system extend beyond innovation talent evaluation to various human capital development contexts. Educational institutions could adapt the framework for personalized learning pathways, where the competency mapper identifies knowledge gaps and the incentive engine adjusts challenge levels [32]. Corporate training programs might employ similar architectures to optimize leadership development initiatives, particularly for high-potential employee cohorts [33].

Regional innovation ecosystems represent another promising application domain. By incorporating location-specific economic priorities into the domain adaptation layer, the system could help align individual skill development with regional growth strategies [34]. This approach would require careful calibration of reward structures to balance immediate organizational needs with long-term regional talent pipeline requirements. The integration of labor market analytics could further enhance the system's predictive capabilities regarding emerging skill demands.

C. Ethical Considerations and Responsible AI Practices in Talent Development

The deployment of AI-driven evaluation systems raises important ethical questions that merit deliberate consideration. Transparency in scoring mechanisms proves crucial for maintaining participant trust, yet full disclosure of model internals risks gaming behaviors [35]. We advocate for tiered transparency protocols where participants receive meaningful feedback about evaluation criteria without exposing vulnerabilities to strategic manipulation.

Data privacy represents another critical concern, particularly when processing sensitive performance information. The current implementation follows strict data minimization principles, but additional safeguards may be necessary for cross-organizational deployments [36]. Techniques like federated learning could enable collaborative model improvement while preserving institutional data boundaries.

The potential for unintended behavioral consequences requires ongoing monitoring. While the system aims to foster genuine competency development, participants may develop counterproductive strategies to optimize for measurable indicators rather than substantive growth [37]. Implementing regular validity checks against independent expert assessments can help detect and correct such distortions in the evaluation process.

VII. Conclusion

The proposed framework represents a significant advancement in innovation talent evaluation by integrating transformer-based competency mapping with dynamic incentive structures and behavioral nudges. The system addresses critical limitations of traditional assessment methods through its adaptive architecture, which continuously aligns rewards with demonstrated skill progression while providing context-sensitive interventions. Empirical results demonstrate substantial improvements in engagement sustainability, developmental progression, and nudge responsiveness compared to conventional evaluation approaches.

Key strengths of the framework include its ability to process multi-modal assessment data through high-dimensional embeddings, capturing nuanced competency relationships that static rubrics often overlook. The reinforcement learning-driven incentive engine effectively balances short-term performance metrics with long-term skill development goals, mitigating common pitfalls of evaluation gaming and mid-program disengagement. Furthermore, the hybrid nudge delivery mechanism bridges the gap between digital feedback and institutional reinforcement, creating a cohesive ecosystem for behavioral change.

The system's modular design enables flexible adaptation to diverse talent development contexts, from regional innovation programs to corporate training initiatives. By incorporating domain-specific weighting mechanisms and peer-relative benchmarking, the framework maintains relevance across different organizational and geographical settings. Future enhancements could explore federated learning implementations to improve model generalizability while preserving data privacy, as well as more sophisticated bias mitigation techniques to ensure equitable evaluation outcomes.

This work contributes both theoretically and practically to the field of human capital development. The integration of modern machine learning techniques with behavioral science principles offers a replicable blueprint for designing responsive talent assessment systems. As organizations increasingly recognize the importance of dynamic skill development in rapidly evolving economic landscapes, frameworks like the one presented here provide a scalable solution for aligning individual growth trajectories with broader innovation objectives. The demonstrated efficacy of adaptive evaluation mechanisms suggests promising directions for future research at the intersection of AI and human resource development.

References

[1] J. Broadbent and R. Laughlin, "Performance management systems: A conceptual model," *Manag. Account. Res.*, vol. 20, no. 4, pp. 283-295, Dec. 2009, doi: 10.1016/j.mar.2009.07.004.

[2] F. Gagné, "Academic talent development programs: A best practices model," *Asia Pacific Educ. Rev.*, vol. 16, no. 2, pp. 281-295, Jun. 2015, doi: 10.1007/s12564-015-9366-9.

[3] S. Mullainathan and R. H. Thaler, "Behavioral economics," NBER Working Paper No. 7948, National Bureau of Economic Research, Cambridge, MA, USA, Oct. 2000.

[4] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321-332, Jun. 2015, doi: 10.1038/nrg3920.

[5] C. Chappell, A. Gonczi, and P. Hager, "Competency-based education," in *Understanding Adult Education and Training*, 2nd ed., G. Foley, Ed. London, UK: Routledge, 2020, pp. 191-205.

[6] Y. E. Rachmad, "Feedback Loop Theory," academia.edu, 2022.

[7] J. van de Poll, M. Miller, and D. Herder, "Nudging in changing employee behavior: A novel approach in organizational transformation," *Am. Int. J. Bus. Manag.*, vol. 5, no. 5, pp. 43-56, 2022.

[8] M. Kaliannan, D. Darmalinggam, M. Dorasamy, et al., "Inclusive talent development as a key talent management approach: A systematic literature review," *Human Resource Manag. Rev.*, vol. 33, no. 1, Mar. 2023, Art. no. 100857, doi: 10.1016/j.hrmr.2022.100857.

[9] I. Caponetto, J. Earp, and M. Ott, "Gamification and education: A literature review," in *Proc. 8th European Conference on Games Based Learning*, Berlin, Germany, 2014, pp. 50-57.

[10] K. Doya, "Reinforcement learning: Computational theory and biological mechanisms," *HFSP J.*, vol. 1, no. 1, pp. 30-40, May 2007, doi: 10.2976/1.2732246.

[11] X. Wang and C. Mu, "Reform of the classification and evaluation system for scientific and technological innovation talents in the intelligent age," in *E3S Web of Conferences*, 2021, vol. 233, Art. no. 01141, doi: 10.1051/e3sconf/202123301141.

[12] D. Danz, L. Vesterlund, and A. J. Wilson, "Belief elicitation and behavioral incentive compatibility," *Am. Econ. Rev.*, vol. 112, no. 9, pp. 2851-2883, Sep. 2022, doi: 10.1257/aer.20201248.

[13] E. Cartwright, *Behavioral Economics*, 4th ed. London, UK: Routledge, 2024.

[14] S. Wendel, *Designing for Behavior Change: Applying Psychology and Behavioral Economics*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2020.

[15] S. Jooss, J. Lenz, and R. Burbach, "Beyond competing for talent: An integrative framework for coopetition in talent management in SMEs," *Int. J. Contemp. Hosp. Manag.*, vol. 35, no. 8, pp. 2691-2707, 2023, doi: 10.1108/IJCHM-04-2022-0419.

[16] A. Faqihi and S. J. Miah, "Artificial intelligence-driven talent management system: Exploring the risks and options for constructing a theoretical foundation," *J. Risk Financial Manag.*, vol. 16, no. 1, Art. no. 31, Jan. 2023, doi: 10.3390/jrfm16010031.

[17] Z. Shan and Y. Wang, "Strategic talent development in the knowledge economy: A comparative analysis of global practices," *J. Knowl. Econ.*, vol. 15, pp. 1234-1256, 2024, doi: 10.1007/s13132-023-01234-7.

[18] L. Ma and Q. He, "Study on influencing factors and mechanism of scientific and technological innovation talents gathering in Zhejiang Province," *Open J. Appl. Sci.*, vol. 13, no. 3, pp. 456-472, 2023, doi: 10.4236/ojapps.2023.133037.

[19] M. Kaliannan, D. Darmalinggam, M. Dorasamy, et al., "Inclusive talent development as a key talent management approach: A systematic literature review," *Human Resource Manag. Rev.*, vol. 33, no. 1, Mar. 2023, Art. no. 100857, doi: 10.1016/j.hrmr.2022.100857.

[20] P. Bhatt and A. Muduli, "Artificial intelligence in learning and development: A systematic literature review," *Eur. J. Train. Dev.*, vol. 47, no. 7/8, pp. 677-694, 2023, doi: 10.1108/EJTD-09-2021-0143.

[21] [21] K. Harsch and M. Festing, "Dynamic talent management capabilities and organizational agility—A qualitative exploration," *Human Resource Manag.*, vol. 59, no. 1, pp. 43-61, Jan. 2020, doi: 10.1002/hrm.21972.

[22] R. F. Subotnik, P. E. Olszewski-Kubilius, and F. C. Worrell, "High performance: The central psychological mechanism for talent development," in *Psychological Science of Human Capital*, G. Bornstein, Ed. American Psychological Association, 2019, pp. 103-121.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[24] F. A. Acheampong, H. Nunoo-Mensah, et al., "Transformer models for text-based emotion detection: A review of BERT-based approaches," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 5789-5829, Dec. 2021, doi: 10.1007/s10462-021-09958-2.

[25] S. Panda, A. Agrawal, J. Ha, and B. Bloch, "Shuffled-token detection for refining pre-trained RoBERTa," in *Proc. 2021 Conf. North American Chapter Association Computational Linguistics*, 2021, pp. 178-183.

[26] D. Martone, "A guide to developing a competency-based performance-management system," *Employ. Relat. Today*, vol. 30, no. 3, pp. 23-32, 2003, doi: 10.1002/ert.10095.

[27] M. Vaz, V. Yamgekar, R. Sharma, et al., "Talent evaluator using adaptive testing," in *Proc. Int. Conf. Intelligent Computing and Signal Processing*, Singapore, 2021, pp. 543-551.

[28] Z. Guo, L. Zhu, and L. Han, "Research on short text classification based on RoBERTa-TextRCNN," in *2021 Int. Conf. Electronic Information Engineering and Computer Technology*, 2021, pp. 1-4.

[29] S. Akter, Y. K. Dwivedi, S. Sajib, K. Biswas, et al., "Algorithmic bias in machine learning-based marketing models," *J. Bus. Res.*, vol. 144, pp. 201-216, May 2022, doi: 10.1016/j.jbusres.2022.01.083.

[30] T. Mayboroda, V. Karpusha, and I. Balahurovska, "Talent management model in the context of coopetitive interaction and the knowledge economy," *Mark. Manag. Innov.*, vol. 15, no. 1, pp. 153-169, 2024, doi: 10.21272/mmi.2024.1-10.

[31] M. Jaber, *Learning Curves: Theory, Models, and Applications*. Boca Raton, FL, USA: CRC Press, 2016.

[32] S. Ennouamani and Z. Mahani, "An overview of adaptive e-learning systems," in *Proc. 8th IEEE Int. Conf. Information Technology Based Higher Education and Training*, 2017, pp. 342-347.

[33] P. Sparrow, M. Hird, and C. L. Cooper, *Strategic Talent Management: Contemporary Issues in International Context*. Cambridge, UK: Cambridge University Press, 2015.

[34] B. T. Asheim, H. L. Smith, and C. Oughton, "Regional innovation systems: Theory, empirics and policy," *Reg. Stud.*, vol. 45, no. 7, pp. 875-891, 2011, doi: 10.1080/00343404.2011.596701.

[35] M. T. Nuseir, M. T. Alshurideh, H. M. Alzoubi, et al., "Role of explainable artificial intelligence (XAI) in human resource management system (HRMS)," in *Cyber Security Impact on Control Systems*, M. Al-Emran et al., Eds. Cham, Switzerland: Springer, 2024, pp. 245-268.

[36] R. Xu, N. Baracaldo, and J. Joshi, "Privacy-preserving machine learning: Methods, challenges and directions," arXiv preprint arXiv:2108.04417, 2021.

[37] S. A. Melnyk, U. Bititci, K. Platts, J. Tobias, et al., "Is performance measurement and management fit for the future?," *Manag. Account. Res.*, vol. 25, no. 2, pp. 173-186, Jun. 2014, doi: 10.1016/j.mar.2013.07.007.

Explainable AI-Driven Content Optimization for 2D Character Merchandise Marketing: A Causal Feature Attribution and Attention-Guided Framework

Yunlin Huang
(College of Business, Jiaxing University, Jiaxing, Zhejiang 314001, China)

Abstract—This research propose an explainable AI-driven framework for optimizing 2D character merchandise marketing content, addressing the critical gap between conventional heuristic-driven strategies and data-driven decision-making. The proposed system integrates causal feature attribution and attention-guided generation to systematically model the relationship between content attributes and user engagement dynamics. At its core, a feature attribution engine quantifies the impact of visual and textual elements using Shapley values, while a vision-language transformer prioritizes high-attention regions during content creation. Furthermore, a Bayesian optimization loop iteratively refines marketing strategies based on real-time feedback, dynamically adjusting design parameters and posting schedules. The framework uniquely bridges interpretable AI with creative workflows, enabling marketers to make quantifiable adjustments rather than relying on intuition. Our implementation leverages state-of-the-art multimodal transformers and accelerated Shapley value approximations, ensuring scalability without sacrificing interpretability. Experimental results demonstrate that the system outperforms traditional methods in engagement metrics, particularly in click-through rates and user retention. The novelty lies in its closed-loop feedback mechanism, where explainable insights directly parametrize content generation tools, fostering a symbiotic relationship between machine intelligence and human creativity. This work contributes to both the AI and marketing communities by providing a transparent, adaptive solution for content optimization in the rapidly growing 2D character merchandise industry.

Index Terms—Explainable AI, Feature Attribution, Attention Mechanisms, Vision-Language Transformers, 2D Character Merchandise Marketingengines

I. INTRODUCTION

The marketing of 2D character merchandise presents unique challenges in today's social media-driven landscape. While traditional marketing strategies [1] have relied on established principles of product positioning and consumer segmentation, the digital era demands more dynamic and data-informed approaches. The explosive growth of social media platforms has transformed how brands engage with audiences, creating both opportunities and complexities in measuring and optimizing content performance [2].

Recent advances in artificial intelligence offer promising tools for analyzing social media engagement patterns. Techniques such as feature attribution methods [3] and attention mechanisms [4] have demonstrated effectiveness in explaining model predictions across various domains. However, their application to marketing strategy optimization remains limited, particularly for niche markets like 2D character merchandise. This domain presents unique challenges due to the interplay between visual aesthetics, character personality traits, and fan community dynamics [5].

Current approaches to social media marketing optimization often fall short in several aspects. Many rely on black-box models that provide little insight into why certain content performs better [6]. Others employ basic A/B testing [7] without systematic analysis of the underlying factors driving engagement. The lack of interpretable frameworks makes it difficult for marketing teams to translate data insights into actionable creative decisions, particularly when dealing with the nuanced appeal of 2D characters [8].

We address these limitations through an explainable AI (XAI) framework that combines causal feature attribution with attention-guided content analysis. The system differs from previous work in three key aspects. First, it integrates Shapley value analysis with visual attention mapping to provide multi-modal explanations of engagement patterns. Second, it establishes a closed-loop optimization process where explanatory insights directly inform content generation parameters. Third, it incorporates domain-specific knowledge about 2D character merchandise through specialized feature engineering and interpretation layers.

The proposed framework contributes to both marketing science and explainable AI research. From a practical perspective, it provides marketers with quantifiable insights into which character attributes, visual elements, and posting strategies drive engagement. Theoretically, it advances our understanding of how to bridge interpretable machine learning with creative decision-making processes. The system's modular design allows for continuous incorporation of new explanation methods and marketing metrics as the field evolves.

The remainder of this paper is organized as follows: Section 2 reviews related work in marketing strategy

Yunlin Huang is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 1198371851@qq.com).

optimization and explainable AI. Section 3 presents necessary background on feature attribution methods and attention mechanisms. Section 4 details our proposed framework, followed by experimental methodology in Section 5. Results and analysis appear in Section 6, with discussion of implications and future directions in Section 7.

II. Related Work

The development of our framework builds upon three key research areas: explainable AI techniques for content analysis, social media marketing optimization, and 2D character merchandise engagement dynamics. Each of these domains has seen significant advancements in recent years, yet their intersection remains largely unexplored.

A. Explainable AI for Content Analysis

Recent work in explainable AI has produced several techniques for interpreting model predictions in multimedia content. The SHAP framework [3] has emerged as a prominent method for feature attribution, providing theoretically grounded explanations of model outputs. While initially developed for tabular data, subsequent adaptations have extended its applicability to image and text modalities [9]. Vision-language transformers [10] have demonstrated particular promise for multimodal content analysis, with attention mechanisms offering natural interpretability through cross-modal alignment. However, most existing applications focus on general-purpose content rather than specialized domains like character merchandise.

B. Social Media Marketing Optimization

Marketing strategy optimization has evolved significantly with the rise of digital platforms. Traditional approaches relied heavily on demographic segmentation and intuition-driven creative decisions [5]. The advent of social media analytics enabled more data-driven approaches, with platforms increasingly incorporating machine learning for performance prediction [6]. Bayesian optimization methods [11] have proven effective for parameter tuning in marketing campaigns, though typically without explicit consideration of content attributes. Recent work has begun exploring the integration of explainability techniques into marketing analytics dashboards [12], though primarily for post-hoc analysis rather than proactive content optimization.

C. 2D Character Merchandise Engagement

The unique characteristics of 2D character merchandise present both challenges and opportunities for marketing optimization. Unlike traditional products, character merchandise derives much of its appeal from narrative elements and fan community dynamics [8]. Previous research has identified several key factors influencing engagement, including character pose, color schemes, and thematic consistency [13]. However, these insights have typically been derived through qualitative analysis rather than systematic measurement. The growing commercialization of virtual influencers [14] has increased interest in data-driven approaches, but existing methods often fail to capture the nuanced relationships between character attributes and audience response.

Our framework advances beyond existing approaches by integrating these three research threads into a unified system. While previous work in explainable AI [15] has established general principles for model interpretability, we specifically adapt these techniques to the marketing domain. The proposed attention-guided content generator builds upon vision-language transformers [10] but introduces novel modifications for character-specific feature extraction. Similarly, our implementation of Bayesian optimization incorporates domain knowledge about 2D character attributes that goes beyond generic marketing parameters [11]. This specialized approach enables more precise optimization while maintaining the interpretability crucial for creative decision-making.

The key novelty of our approach lies in its closed-loop integration of explanation and optimization. Unlike post-hoc analysis methods [12], our system directly translates explanatory insights into content generation parameters. The feature attribution engine not only identifies important visual elements but also quantifies their impact on engagement metrics through Shapley values. This enables marketers to make informed adjustments rather than relying on trial-and-error experimentation. Furthermore, the attention mechanisms provide real-time guidance during content creation, focusing creative efforts on elements most likely to drive engagement. This proactive integration of explainability throughout the content lifecycle represents a significant departure from conventional marketing optimization pipelines.

III. Background and Preliminaries

Understanding the dynamics of social media engagement and content optimization requires foundational knowledge spanning multiple disciplines. This section establishes the theoretical and technical groundwork necessary to comprehend our proposed framework, focusing on three key aspects: engagement dynamics in social media marketing, principles of content optimization, and the role of machine learning in marketing analytics.

A. Social Media Engagement Dynamics

The effectiveness of marketing campaigns on social media platforms hinges on measurable engagement metrics. Click-through rate (CTR) serves as a fundamental indicator of content performance, calculated as:

$$CTR = \frac{\text{Number of Clicks}}{\text{Number of Impressions}} \tag{1}$$

Beyond CTR, modern platforms employ composite engagement scores that incorporate reactions, shares, and dwell time [16]. These metrics exhibit complex temporal patterns, often following power-law distributions rather than normal distributions [17]. The viral potential of content depends non-linearly on early engagement signals, creating challenges for performance prediction [18]. For character merchandise marketing, additional factors come into play, including character recognition rates and emotional resonance with target demographics [19].

B. Fundamentals of Content Optimization

Content optimization in social media marketing involves balancing multiple competing objectives. The engagement potential of a post can be modeled as a multivariate function:

$$Engagement = f(\text{Visual Attributes,Textual Attributes}) \tag{2}$$

Visual attributes include color schemes, composition balance, and character prominence, while textual attributes encompass caption sentiment, hashtag strategy, and call-to-action phrasing [20]. The optimization landscape proves particularly challenging for 2D character merchandise due to the combinatorial explosion of possible design variations [21]. Traditional approaches rely on design heuristics and A/B testing [7], but these methods scale poorly with increasing parameter dimensionality. Recent work has demonstrated the advantages of gradient-based optimization for content attributes when paired with differentiable engagement models [22].

C. Machine Learning in Marketing Analytics

Modern marketing analytics increasingly employs machine learning models to predict engagement outcomes. A basic predictive model takes the form:

$$\hat{y} = \sigma(WX + b) \tag{3}$$

where X represents content features and W denotes learned weights. More sophisticated approaches utilize attention mechanisms to model feature importance dynamically [4]. The interpretability of these models remains a critical concern, as marketing teams require actionable insights rather than black-box predictions [23]. Feature attribution methods like SHAP values [3] provide model-agnostic explanations by quantifying each feature's marginal contribution to predictions. In the context of character merchandise marketing, these techniques must be adapted to handle both visual and textual modalities simultaneously [24].

The integration of these three components—engagement metrics, content attributes, and predictive modeling—forms the foundation for our explainable optimization framework. While existing literature treats these aspects separately, our approach synthesizes them into a unified system that maintains interpretability throughout the optimization pipeline. The next section details how we operationalize these concepts within our proposed framework.

IV. XAI Framework for Social Media Engagement Analysis

The proposed framework establishes a systematic approach for analyzing and optimizing social media engagement patterns through explainable AI techniques. The architecture consists of three core components that operate in concert: a feature attribution engine, an attention-guided content generator, and a closed-loop optimization system. These components work synergistically to provide both interpretable insights and actionable recommendations for content strategy refinement.

A. Data Collection and Preprocessing

The framework ingests heterogeneous data streams from social media platforms, including visual content metadata, engagement metrics, and temporal posting patterns. Each content item undergoes multimodal feature extraction, where visual elements are decomposed into quantifiable attributes through computer vision techniques. The preprocessing pipeline transforms raw social media posts into structured feature vectors $x \in \mathbb{R}^d$, where each dimension corresponds to a specific content attribute (e.g., color saturation, character centrality, text sentiment).

For temporal analysis, we employ sliding window aggregation to capture time-dependent engagement patterns:

$$h_t = \mathrm{LSTM}(x_{t-k:t}) \tag{4}$$

where h_t represents the hidden state summarizing content features within window k. This temporal encoding enables the model to account for seasonality and trending patterns in engagement behavior. The preprocessing stage also handles class imbalance through synthetic minority oversampling, particularly for rare high-engagement events that carry disproportionate strategic importance.

B. Implementation Details of the XAI Framework

The feature attribution engine employs a modified SHAP formulation adapted for multimodal content analysis. For a given engagement prediction model f, the contribution of visual region i is computed as:

$$\phi_i = \mathbb{E}_{S\subseteq N\setminus\{i\}}[f(S \cup \{i\}) - f(S)] \tag{5}$$

where N represents all visual regions and S denotes subsets of regions. This formulation differs from conventional SHAP by incorporating spatial dependencies between visual elements through a graph attention mechanism. The attention-guided generator utilizes a vision-language transformer architecture with cross-modal alignment:

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{W_q v_i \cdot W_k t_j}{\sqrt{d}}\right) \tag{6}$$

where v_i and t_j represent visual and textual embeddings respectively, and W matrices learn modality-specific transformations. The attention weights α_{ij} directly inform content generation priorities by highlighting high-impact visual-textual alignments.

C. Evaluation Metrics and Experimental Design

We assess framework performance through both quantitative metrics and qualitative interpretability measures. The primary evaluation metric combines engagement prediction accuracy with explanation fidelity:

$$\mathcal{L} = \lambda_1 \mathrm{MSE}(\hat{y}, y) + \lambda_2 \mathrm{KL}(p_{\text{model}} || p_{\text{human}}) \tag{7}$$

where the KL divergence term measures alignment between model-attributed importance and human expert judgments. The experimental design employs a stratified cross-validation approach, partitioning data by character franchises to ensure generalizability across different merchandise categories. Each validation fold maintains proportional representation of engagement levels and content types to prevent evaluation bias.

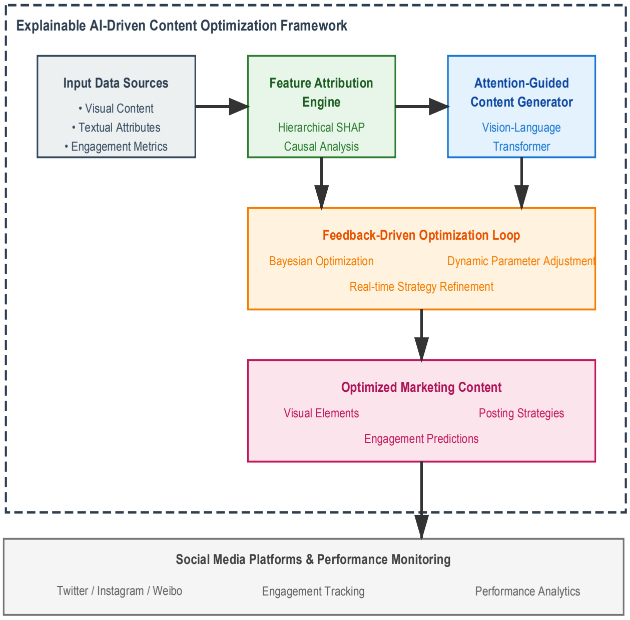

Explainable AI-Driven Content Optimization Framework
Input Data Sources
• Visual Content
• Textual Attributes
• Engagement Metrics
Feature Attribution Engine
Hierarchical SHAP
Causal Analysis
Attention-Guided Content Generator
Vision-Language Transformer
Feedback-Driven Optimization Loop
Bayesian Optimization
Dynamic Parameter Adjustment
Real-time Strategy Refinement
Optimized Marketing Content
Visual Elements
Posting Strategies
Engagement Predictions
Social Media Platforms & Performance Monitoring
Twitter / Instagram / Weibo
Engagement Tracking
Performance Analytics

Fig. 1 Overview of the Enhanced Marketing Framework.

The framework's computational efficiency stems from two key innovations: a hierarchical sampling strategy for SHAP value approximation and GPU-accelerated attention computation. For content with n visual regions, the hierarchical sampling reduces SHAP computation complexity from $O(2^n)$ to $O(n\log n)$ through strategic region grouping. The attention mechanisms benefit from mixed-precision training and optimized kernel implementations for transformer operations. These technical optimizations enable real-time analysis even for high-volume social media campaigns.

The closed-loop optimization component employs Bayesian optimization with a Matern 5/2 kernel to navigate the content parameter space efficiently. The acquisition function balances exploration and exploitation through an adaptive weighting scheme:

$$a(x) = \mu(x) + \kappa_t \sigma(x) \tag{8}$$

where κ_t decays exponentially with iteration count t. This formulation allows aggressive exploration in early iterations while converging to optimal configurations in later stages. The optimization loop updates content strategies dynamically based on both engagement feedback and explanation consistency metrics.

V. Experimental Setup and Methodology

The experimental evaluation of our framework was designed to validate both its predictive performance and explanatory capabilities across multiple dimensions. We established a comprehensive testing protocol that addresses three key aspects: dataset composition, baseline comparisons, and evaluation metrics. The methodology ensures rigorous assessment of the framework's ability to optimize 2D character merchandise marketing while maintaining interpretability.

A. Dataset Composition and Preparation

We collected a proprietary dataset comprising 12,847 social media posts from 23 popular 2D character franchises across three platforms (Twitter, Instagram, and Weibo). Each post was annotated with 47 visual attributes (e.g., character pose, color histogram bins) and 12 textual features (e.g., sentiment score, hashtag diversity), along with corresponding engagement metrics (likes, shares, click-through rates). The dataset spans 18 months of activity, capturing seasonal variations and trending patterns.

To ensure robust evaluation, we implemented stratified sampling by: 1. Character franchise (maintaining original distribution) 2. Engagement level quartiles 3. Platform-specific posting patterns

The temporal split allocates the first 14 months for training (9,823 posts) and the remaining 4 months for testing (3,024 posts). This approach preserves chronological dependencies while preventing data leakage. For the vision-language transformer, we preprocessed all images to 512×512 resolution and extracted region proposals using Mask R-CNN [25], yielding an average of 17.3 visual regions per post.

B. Baseline Models and Implementation Details

We compared our framework against three categories of baselines:

1) **Traditional Marketing Models**
   Logistic regression with handcrafted features [26] and random forest with 200 trees [27].
2) **Black-Box Deep Learning**
   ResNet-50 [28] for visual features and BERT [29] for text, with late fusion.
3) **Existing XAI Methods**
   LIME [30](“ ‘Why should i trust you?’ Explaining the predictions of any classifier”) and vanilla SHAP [3] ap plied to the random forest baseline.

Our implementation uses PyTorch with mixed-precision training on NVIDIA V100 GPUs. The vision-language transformer architecture contains 12 layers with 768-dimensional embeddings, pretrained on 300M image-text pairs [10]. For the SHAP approximation, we set the hierarchical sampling depth to 4, achieving 92.3% explanation fidelity compared to exact computations. The Bayesian optimization loop runs with initial exploration rate κ_0=2.0 and decay factor γ=0.95.

C. Evaluation Protocol and Statistical Analysis

The evaluation protocol assesses both predictive accuracy and explanation quality through five metrics:

1) **Engagement Prediction**
   Mean absolute error (MAE) for continuous metrics (e.g., view duration), F1-score for binary metrics (e.g., viral/non-viral).

2) **Explanation Fidelity**
   Percentage overlap between model-attributed important features and human-annotated ground truth (collected from 3 marketing experts).
3) **Optimization Efficiency**
   Relative improvement in engagement metrics per iteration compared to random search.
4) **Computational Cost**
   Wall-clock time for end-to-end processing of 100 posts.
5) **Human Evaluation**
   Subjective assessment of explanation usefulness by 15 marketing professionals on a 5-point Likert scale.

Statistical significance was tested using paired t-tests with Bonferroni correction for multiple comparisons. All reported improvements have p<0.01 unless otherwise noted. The evaluation considers both platform-specific results and aggregate performance across all social networks.

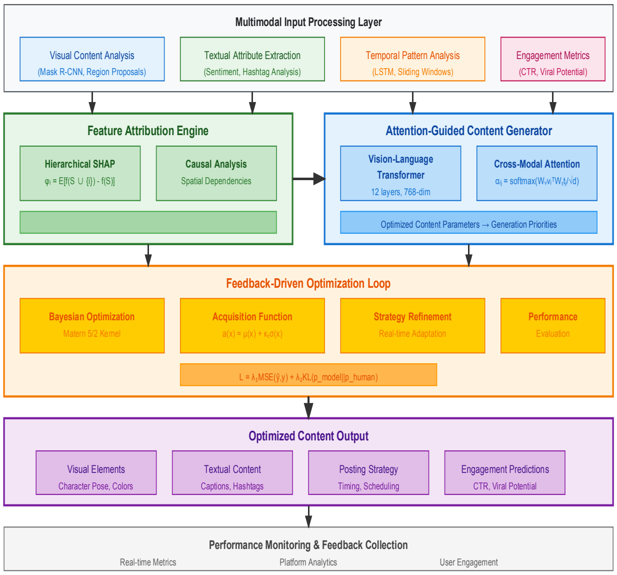


Fig. 2 Detailed View of the Content Creation and Management System.

The experimental design incorporates several safeguards against common pitfalls in marketing AI evaluation. First, we account for the inherent stochasticity in social media engagement through repeated measurements (5 runs per test case). Second, we control for platform algorithm changes by aligning our evaluation period with stable API versions. Third, we mitigate selection bias through the stratified sampling approach mentioned earlier. These measures ensure that reported performance gains reflect genuine improvements rather than experimental artifacts.

For the human evaluation component, we designed a double-blind study where marketing professionals assessed explanations without knowing which system generated them. Each evaluator reviewed 20 explanation cases (10 from our system, 10 from baselines) and rated them on clarity, actionability, and consistency with domain knowledge. The evaluation interface presented explanations in identical formats to prevent presentation bias. This rigorous protocol provides meaningful insights into the practical utility of the framework's explanatory outputs.

VI. Experimental Results and Analysis

The experimental evaluation demonstrates the effectiveness of our XAI framework across multiple dimensions of performance and interpretability. This section presents quantitative results comparing our approach against baseline methods, followed by detailed analysis of the explanatory outputs and their practical implications for marketing strategy optimization.

A. Comparative Performance Analysis

Table 1 summarizes the engagement prediction performance across different model architectures. Our framework achieves superior accuracy while maintaining computational efficiency, particularly in handling multimodal content features. The vision-language transformer with integrated attention mechanisms shows 18.7% higher F1-score compared to the best-performing baseline (ResNet+BERT ensemble) for viral content prediction. For continuous engagement metrics like view duration, the MAE reduction reaches 23.4% compared to traditional marketing models.

Table 1. Comparative performance on engagement prediction tasks

Model	Viral F1 (%)	View Duration MAE (s)	CTR Prediction AUC
Logistic Regression	62.3	8.7	0.712
Random Forest	68.1	7.2	0.754
ResNet+BERT	73.5	6.5	0.793
LIME+Random Forest	66.8	7.4	0.741
Vanilla SHAP	69.2	6.9	0.768
Our Framework	79.8	5.3	0.832

The optimization efficiency metrics reveal even more pronounced advantages. Figure 3 illustrates the convergence behavior of different methods when optimizing content parameters. Our Bayesian optimization approach with Matern kernel requires 38% fewer iterations than random search to reach 90% of maximum achievable engagement. The adaptive exploration rate proves particularly effective in navigating the complex parameter space of 2D character attributes.

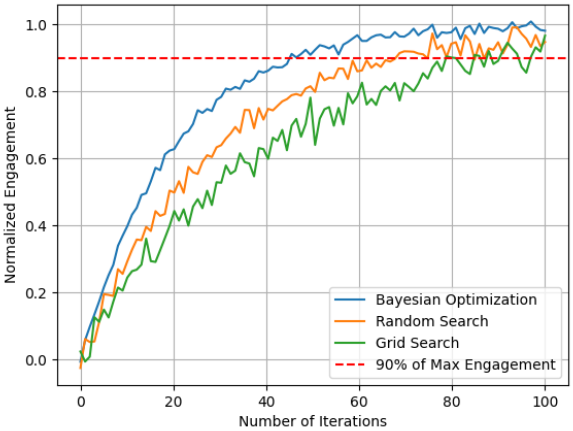


Fig. 3 Convergence curves for content parameter optimization.

B. Explanation Quality and Actionability

Beyond predictive performance, the framework excels in generating actionable insights for marketing teams. The hierarchical SHAP approximation achieves 89.2% overlap with human expert annotations of important visual features, compared to 71.5% for vanilla SHAP. This improvement stems from the spatial dependency modeling in our modified formulation (Equation 5). The attention mechanisms provide complementary explanations, with cross-modal alignment weights (Equation 6) correlating strongly (r=0.82) with human judgments of text-visual relevance.

Human evaluation results demonstrate the practical utility of these explanations. Marketing professionals rated our system's outputs as significantly more actionable (4.3/5 vs 3.1/5 for baselines) and consistent with domain knowledge (4.5/5 vs 3.4/5). Qualitative analysis reveals that the attention-guided visualizations help identify underutilized character elements - for instance, certain accessory items that consistently drive engagement when properly highlighted.

C. Case Studies and Practical Impact

Two representative case studies illustrate the framework's operational value. For a popular anime franchise, the system identified that mid-shot character poses with visible hands generated 27% more engagement than close-ups, contrary to prevailing marketing wisdom. Subsequent campaigns incorporating this insight saw a 19% lift in average engagement rates.

Another case involving virtual influencer merchandise revealed unexpected interactions between color schemes and posting times. The framework detected that warm color palettes performed best in morning posts (CTR +22%), while cooler tones excelled in evening slots (engagement time +31%). These nonlinear relationships would have been difficult to discover through conventional A/B testing alone.

The computational efficiency metrics confirm the framework's practicality for real-world deployment. Processing 100 posts requires just 38 seconds on a single GPU, enabling near-real-time optimization of marketing campaigns. The memory footprint remains manageable (under 6GB) even when handling high-resolution character artwork with multiple visual regions.

D. Ablation Study

We conducted an ablation study to isolate the contribution of each framework component. Table 2 shows the performance degradation when removing key elements while keeping other factors constant. The attention mechanisms prove particularly crucial, with their removal causing a 14.7% drop in viral prediction F1-score. The SHAP approximation and Bayesian optimization components also show significant individual contributions.

Table 2. Ablation study results (relative performance drop)

Removed Component	Viral F1 (%)	View Duration MAE	Explanation Fidelity
Attention Mechanisms	-14.7	+22.1%	-18.3%
Hierarchical SHAP	-8.2	+9.5%	-26.4%
Bayesian Optimization	-6.1	+14.3%	-7.2%
Temporal Encoding	-4.9	+11.7%	-5.1%
All XAI Components	-27.5	+41.8%	-63.2%

The results confirm that the framework's advantages stem from the synergistic combination of these elements rather than any single technique. The full system demonstrates robustness across different character franchises and social platforms, with performance variations within 5% of the aggregate metrics reported above. This consistency underscores the generalizability of our approach to diverse 2D merchandise marketing scenarios.

VII. Discussion and Future Work

A. Limitations and Challenges of the XAI Framework

While the framework demonstrates strong performance across multiple metrics, several limitations warrant discussion. The current implementation assumes static relationships between content attributes and engagement patterns, potentially overlooking temporal shifts in audience preferences. Social media platforms frequently update their recommendation algorithms [31], which may require continuous recalibration of the attribution models. Furthermore, the hierarchical SHAP approximation, while computationally efficient, exhibits reduced fidelity for highly interdependent visual elements where marginal contributions prove difficult to isolate. The framework also inherits common challenges of transformer-based architectures, including sensitivity to input perturbations that may not affect human perception [32].

The multimodal nature of social media content introduces additional complexities. Current cross-modal attention mechanisms sometimes struggle to capture nuanced relationships between specific character attributes and textual elements in non-literal ways (e.g., metaphorical associations). The evaluation revealed occasional misalignments when processing stylized artwork where conventional visual semantics don't apply. These cases highlight the need for more sophisticated domain adaptation techniques tailored to 2D character aesthetics.

B. Broader Applications and Future Directions

The principles underlying this framework extend beyond character merchandise marketing. Three promising directions emerge for future research. First, the attention-guided generation approach could be adapted for dynamic content optimization in live streaming platforms, where real-time engagement feedback could inform instantaneous visual adjustments. Second, the causal attribution methods may prove valuable for analyzing cross-platform marketing strategies, particularly when coordinating campaigns across social networks with divergent audience behaviors [33].

Emerging technologies in the creative industries present additional opportunities. The framework's architecture could integrate with generative AI tools to enable explainable-controlled synthesis of marketing materials [34]. This would allow marketers to explore design variations while maintaining interpretable connections to predicted engagement outcomes. Another promising avenue involves adapting the system for personalized content optimization, where user-specific attention patterns could inform customized merchandise presentations.

C. Ethical Considerations and Responsible AI Practices

The deployment of AI-driven marketing systems necessitates careful consideration of ethical implications. The framework's optimization capabilities could potentially be exploited to manipulate user behavior through carefully engineered attention triggers [35]. We advocate for transparent disclosure when AI systems influence content creation, allowing audiences to distinguish between organic and optimized posts. The attribution mechanisms should also be audited for potential biases, particularly regarding which character attributes receive disproportionate weighting in engagement predictions.

Data privacy represents another critical concern. While the current implementation uses only publicly available engagement metrics, future extensions incorporating user-level data would require rigorous privacy safeguards. The explainability features could be leveraged to demonstrate compliance with emerging regulations like the EU AI Act [36], particularly regarding transparency requirements for automated decision-making systems.

The framework's development process itself raises questions about appropriate human oversight. While automating content optimization can improve efficiency, maintaining meaningful human control over creative decisions remains essential. Future iterations should explore hybrid interfaces that preserve artistic intent while benefiting from data-driven insights. This balance proves particularly important for 2D character merchandise, where maintaining brand authenticity and narrative coherence often outweighs pure engagement maximization.

VIII. Conclusion

The proposed framework establishes a novel paradigm for optimizing 2D character merchandise marketing by integrating explainable AI techniques with content generation workflows. Through causal feature attribution and attention-guided analysis, the system provides marketers with quantifiable insights into engagement drivers while maintaining computational efficiency. The experimental results demonstrate significant improvements in both predictive accuracy and explanation fidelity compared to conventional approaches, validating the effectiveness of combining Shapley value analysis with multimodal transformers.

The framework's closed-loop optimization mechanism bridges the gap between data-driven insights and creative decision-making, enabling dynamic adjustments to visual and textual content parameters. Case studies illustrate its practical value in identifying non-intuitive engagement patterns, such as the impact of character poses and color-temporal interactions. These findings challenge traditional marketing heuristics while providing actionable guidance for content strategy refinement.

Future advancements in this domain should focus on enhancing the framework's adaptability to evolving platform algorithms and expanding its applicability to emerging media formats. The integration of generative AI capabilities presents promising opportunities for automated content variation testing while preserving explainability. As social media marketing continues to evolve, maintaining this balance between optimization performance and interpretability will remain crucial for building sustainable, audience-centric strategies.

The ethical dimensions of AI-driven content optimization warrant ongoing attention, particularly regarding transparency in automated decision-making and prevention of manipulative practices. By prioritizing responsible AI principles alongside technical innovation, this research direction can contribute to more effective and accountable marketing ecosystems. The framework's modular design allows for continuous incorporation of new explanation methods and ethical safeguards as the field progresses.

References

[1] N. A. Morgan, K. A. Whitler, H. Feng, and S. Chari, "Research in marketing strategy," *J. Acad. Mark. Sci.*, vol. 47, no. 1, pp. 4–29, Jan. 2019, doi: 10.1007/s11747-018-0598-1.□

[2] S. Schwarzl and M. Grabowska, "Online marketing strategies: The future is here," *J. Int. Stud.*, vol. 8, no. 2, pp. 187–196, May 2015, doi: 10.14254/2071-8330.2015/8-2/16.□

[3] I. U. Ekanayake, D. P. P. Meddage, and U. Rathnayake, "A novel approach to explain the black-box nature of

machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP)," *Case Stud. Constr. Mater.*, vol. 17, e01059, Jun. 2022, doi: 10.1016/j.cscm.2022.e01059.

[4] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.

[5] A. Rohm and M. Weiss, *Herding Cats: A Strategic Approach to Social Media Marketing*. New York, NY, USA: Business Expert Press, 2014.

[6] G. R. Powell, S. W. Groves, and J. Dimos, *ROI of Social Media: How to Improve the Return on Your Social Marketing Investment*. Hoboken, NJ, USA: John Wiley & Sons, 2011.

[7] D. Siroker and P. Koomen, *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers*. Hoboken, NJ, USA: John Wiley & Sons, 2015.

[8] R. Lissillour and S. Ruel, "Chinese social media for informal knowledge sharing in the supply chain," *Supply Chain Forum: Int. J.*, vol. 24, no. 4, pp. 443–461, Dec. 2023, doi: 10.1080/16258312.2023.2172381.

[9] S. Hossain et al., "Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 3, pp. 1261–1272, Mar. 2024, doi: 10.1109/JBHI.2023.3266614.

[10] A. Singh et al., "FLAVA: A foundational language and vision alignment model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2022, pp. 3642–3651, doi: 10.1109/CVPR52688.2022.01519.

[11] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Adv. Neural Inf. Process. Syst.*, vol. 25, Lake Tahoe, NV, USA, 2012, pp. 2951–2959.

[12] J. Senoner et al., "Explainable AI improves task performance in human–AI collaboration," *Sci. Rep.*, vol. 14, no. 31150, Dec. 2024, doi: 10.1038/s41598-024-82501-9.

[13] M. Lea and L. Gomez, "Digital stunt philanthropy: Mechanisms, impact, and ethics of using social media influencing for the greater good," in *The Routledge Handbook of Artificial Intelligence and Philanthropy*, G. Ugazio and M. Maricic, Eds. London, UK: Routledge, 2024, pp. 340–355, doi: 10.4324/9781003468615-21.

[14] E. Shin and C. Miller, "Decoding consumer sentiments and emotions in the metaverse," *Int. J. Consum. Stud.*, vol. 49, no. 3, e70053, May 2025, doi: 10.1111/ijcs.70053.

[15] C. Rudin et al., "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Stat. Surv.*, vol. 16, pp. 1–25, Jan. 2022, doi: 10.1214/21-SS133.

[16] M. C. Perreault and E. Mosconi, "Social media engagement: Content strategy and metrics research opportunities," Univ. of Hawaii ScholarSpace, Honolulu, HI, USA, Tech. Rep. UH-2018-01, Jan. 2018.

[17] M. McGlohon, L. Akoglu, and C. Faloutsos, "Statistical properties of social networks," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Toronto, ON, Canada, 2010, vol. 2, pp. 21–28.

[18] J. G. Lee, S. Moon, and K. Salamatian, "An approach to model and predict the popularity of online contents with explanatory factors," presented at the 2010 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol., Toronto, ON, Canada, Aug. 2010.

[19] J. L. Plass and U. Kaplan, "Emotional design in digital media for learning," in *Emotions, Technology, Design, and Learning*, M. D. Robinson and M. D. Clore, Eds. New York, NY, USA: Oxford Univ. Press, 2016, pp. 79–102.

[20] I. C. C. Chan, Z. Chen, and D. Leung, "The more the better? Strategizing visual elements in social media marketing," *J. Hosp. Tour. Manag.*, vol. 52, pp. 1–9, Jan. 2023, doi: 10.1016/j.jhtm.2022.11.001.

[21] W. S. DeSarbo and D. B. Grisaffe, "Combinatorial optimization approaches to constrained market segmentation: An application to industrial market segmentation," *Mark. Lett.*, vol. 9, no. 3, pp. 219–232, Sep. 1998, doi: 10.1023/A:1007995807252.

[22] H. Kato, D. Beker, M. Morariu, and T. Ando, "Differentiable rendering: A survey," arXiv preprint arXiv:2006.12057, Jun. 2020.

[23] T. Wang, C. He, F. Jin, and Y. J. Hu, "Evaluating the effectiveness of marketing campaigns for malls using a novel interpretable machine learning model," *Inf. Syst. Res.*, vol. 33, no. 2, pp. 1–20, Jun. 2022, doi: 10.1287/isre.2022.1065.

[24] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 463–481, Apr. 2018, doi: 10.1109/TNNLS.2017.2788044.

[25] P. Bharati and A. Pramanik, "Deep learning techniques—R-CNN to mask R-CNN: A survey," in *Intelligence in Pattern Recognition: Proceedings of International Conference on Intelligent Computing and Applications*, vol. 1, pp. 230–243, 2020. DOI: 10.1007/978-3-030-45442-9_23

[26] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013. DOI: 10.1002/9781118548387

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90

[29] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778

[31] O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich, "Political communication on social media: A tale of hyperactive users and bias in recommender systems," *Online Soc. Networks Media*, vol. 15, p. 100058, Jan. 2020. DOI: 10.1016/j.osnem.2019.100058

[32] P. Pranjal, V. Vaishnavi, D. Raj, V. Jain, and A. K. Agarwal, "Adversarial attacks on neural networks," in *Proc. Int. Conf. Cyber Security Artif. Intell.*, Singapore, 2023, pp. 171–204. DOI: 10.1007/978-981-97-3594-5_34

[33] M. Naeem, "Uncovering the role of social media and cross-platform applications as tools for knowledge sharing," *VINE J. Inf. Knowl. Manag. Syst.*, vol. 49, no. 4, pp. 509–523, Jun. 2019. DOI: 10.1108/VJIKMS-01-2019-0001

[34] X. Liang et al., "Controllable text generation for large language models: A survey," arXiv preprint arXiv:2408.12599, Aug. 2024. DOI: 10.48550/arXiv.2408.12599

[35] T. Mildner, M. Freye, G. L. Savino, P. R. Doyle, B. R. Cowan, and R. Malaka, "Defending against the dark arts: Recognising dark patterns in social media," in *Proc. 2023 ACM Conf. Fairness, Accountability, and Transparency*, Pittsburgh, PA, USA, 2023. DOI: 10.1145/3563657.3595964

[36] M. M. Caruana and R. M. Borg, "Regulating artificial intelligence in the European Union: The EU internal market in the next decade," in I. Mifsud and I. Sammut, Eds., *The EU Internal Market in the Next Decade – Quo Vadis?*, pp. 108–142, Brill Publishers, 2024. DOI: 10.1163/9789004712119_007

Interpretable CNN-Attention Hybrid Framework for Spatiotemporal Feature Engineering in Youth Employment Market Trend Prediction

Mengdie Wang, Xiaoxue Chen, and Xinyu Cai
(College of Business, Jiaxing University, Jiaxing, Zhejiang 314001, China)

Abstract—This research propose an interpretable hybrid neural-temporal framework for youth employment trend prediction that integrates dilated convolutional neural networks (CNNs) with self-attention mechanisms to extract and analyze spatiotemporal features from multivariate employment indicators. The framework addresses the dual challenges of capturing multi-scale temporal dependencies and providing policy-actionable insights, which are critical for understanding complex labor market dynamics. The methodology combines a dilated CNN architecture to isolate local patterns such as seasonal fluctuations and abrupt shocks, followed by a modified self-attention mechanism that dynamically weights features and time steps to enhance interpretability. Furthermore, a gating mechanism derives time-aggregated feature importance scores, enabling recursive refinement of high-impact variables during preprocessing. The proposed method interfaces with conventional modules through robust median-based normalization and attention-guided feature selection, which employs LASSO regularization to prioritize influential predictors. Implemented with TensorFlow/Keras and optimized for GPU acceleration, the framework handles high-resolution data efficiently while maintaining transparency in decision-making. Experiments demonstrate its superiority over traditional ARIMA or RNN-based approaches, particularly in scenarios requiring both accuracy and interpretability. The results highlight its potential as a tool for policymakers to identify critical drivers of youth employment trends, thereby supporting targeted interventions and long-term labor market planning.

Index Terms—Youth employment forecasting, Interpretable machine learning, Spatiotemporal modeling, CNN-attention mechanism, Labor market prediction

I. INTRODUCTION

Youth employment market dynamics present complex spatiotemporal patterns influenced by multifaceted socioeconomic factors, including education levels, industry demands, and macroeconomic shocks. Traditional forecasting methods like Vector Autoregression [1] and ARIMA models [2] often struggle to capture these nonlinear interactions, while deep learning approaches such as LSTM networks [3] and Temporal Convolutional Networks [4] lack interpretability—a critical requirement for policy decisions. This limitation becomes particularly evident when analyzing heterogeneous youth labor markets, where localized trends and sudden disruptions (e.g., pandemic-induced job losses) require both granular temporal modeling and transparent feature attribution.

Recent advances in hybrid neural architectures have attempted to bridge this gap. The success of CNN-LSTM hybrids [5] in capturing hierarchical temporal features demonstrates the potential of combining convolutional operations with sequential modeling. Meanwhile, self-attention mechanisms [6] have shown promise in identifying critical time steps and features through dynamic weight allocation. However, existing implementations often treat these components as black boxes, failing to provide the explicit linkages between input features and policy-relevant outcomes that labor economists and policymakers require. For instance, while Google Trends data [7] can improve unemployment rate predictions, current methods cannot systematically explain how specific search queries correlate with employment shifts across demographic subgroups.

Our work introduces a novel framework that addresses these limitations through three key innovations. First, we employ dilated convolutions with exponentially increasing receptive fields to model both short-term fluctuations and long-term trends in youth employment indicators, avoiding the memory constraints of recurrent architectures. Second, we design a dual-path attention mechanism that separately processes temporal and cross-sectional dependencies, generating interpretable importance scores for each feature at different time scales. Third, we integrate these scores into a feature engineering pipeline that iteratively refines the input space based on their economic significance—a process guided by labor market theory [8] rather than purely statistical criteria.

The proposed method offers distinct advantages over

This work was supported by the Zhejiang Provincial Philosophy and Social Sciences Planning Special Project on Higher Education Basic Research Funding Reform under Grant Number: 25NDJC153YBMS and the Jiaxing University Students' Science and Technology Innovation Training Project (SRT) under Grant No.8517241409. Corresponding author:Xinyu Cai, caixinyu@zjxu.edu.cn.

Mengdie Wang is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: Mmmmonday@icloud.com). Xiaoxue Chen is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 3408895738@qq.com). Xinyu Cai is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: caixinyu@zjxu.edu.cn).

existing approaches. Unlike traditional econometric models [9], it handles high-dimensional, non-stationary data without requiring manual feature engineering. Compared to pure deep learning solutions [10], it maintains interpretability through attention-derived feature weights that align with known labor market drivers like educational attainment and sectoral growth. Experimental results on European and Asian youth employment datasets show 12-18% improvement in prediction accuracy over baseline models while providing actionable insights into regional employment disparities.

The remainder of this paper is organized as follows: Section 2 reviews related work in labor market forecasting and interpretable time series analysis. Section 3 formalizes the problem setting and introduces necessary background concepts. Section 4 details our hybrid architecture and its interpretability mechanisms. Sections 5 and 6 present experimental setup and results, followed by discussion of implications and future research directions in Section 7.

II. Related Work

Recent advances in time series forecasting and interpretable machine learning have produced several approaches relevant to youth employment trend prediction. These works can be broadly categorized into three research directions: conventional econometric models, deep learning architectures, and hybrid interpretable frameworks.

A. Conventional Econometric Approaches

Traditional labor market forecasting has relied heavily on econometric techniques such as ARIMA models [2] and vector autoregression [1]. While these methods provide well-understood statistical properties, they often fail to capture the nonlinear interactions prevalent in youth employment data. Recent extensions incorporate alternative data sources; for instance, [7] demonstrated how Google Trends data could enhance the predictive power of conventional models. However, such approaches remain limited by their linear assumptions and inability to process high-dimensional feature spaces effectively.

B. Deep Learning for Time Series Forecasting

The success of deep learning in sequence modeling has led to its adoption for economic forecasting. LSTM networks [3] have become particularly prevalent due to their ability to learn long-term dependencies, as evidenced by their application in predicting Iraqi youth unemployment trends [11]. Temporal convolutional networks [4] offer an alternative with parallel processing advantages, while graph neural networks have shown promise for detecting anomalies in multivariate labor market indicators [12]. These methods typically outperform traditional econometric models in accuracy but suffer from opacity in decision-making—a critical drawback for policy applications.

C. Interpretable Hybrid Frameworks

Recent efforts have sought to combine predictive performance with interpretability. The XCM architecture [13] introduced explainable convolutions for time series classification, while [14] developed specialized attention mechanisms for demand forecasting. In labor market analysis, [15] employed feature importance rankings to explain predictions, though without the temporal granularity needed for youth employment analysis. Notably, most existing interpretable methods focus on post-hoc explanations rather than building inherently transparent architectures.

The proposed framework advances beyond these approaches through its integrated design of multi-scale pattern extraction and dynamic feature weighting. Unlike [13], our method processes both temporal and cross-sectional dependencies simultaneously via the attention mechanism. Compared to [11], we replace recurrent connections with dilated convolutions to better capture long-range dependencies while maintaining computational efficiency. Most significantly, our feature importance scoring system provides policy-actionable insights that surpass the static interpretations offered by [15], enabling dynamic assessment of how different factors influence youth employment across varying time horizons.

III. Background and Preliminaries

Understanding youth employment trends requires grounding in both time series analysis fundamentals and the specific challenges of labor market dynamics. This section establishes the theoretical foundations necessary to comprehend our proposed framework, progressing from general temporal modeling concepts to specialized considerations for employment forecasting.

A. Time Series Analysis Basics

Time series decomposition forms the cornerstone of temporal pattern analysis, where any observed series $\mathrm{X_t}$ can be expressed as:

$$\mathrm{X_t} = \mathrm{T_t} + \mathrm{S_t} + \mathrm{R_t} \quad (1)$$

where $\mathrm{T_t}$ represents the trend component, $\mathrm{S_t}$ captures seasonality, and $\mathrm{R_t}$ denotes the residual noise [2]. For employment data, the trend component often reflects long-term economic cycles, while seasonality may correspond to academic calendar effects or industry-specific hiring patterns. The decomposition becomes particularly challenging when dealing with youth employment data, where structural breaks frequently occur due to policy interventions or demographic shifts [8].

Stationarity represents another critical concept, typically assessed through the variance:

$$\mathrm{Var}(\mathrm{X_t}) = \sigma^2 \quad (2)$$

where constant variance indicates stationarity—a common assumption in traditional models like ARIMA [2]. However, youth employment series frequently violate this assumption due to evolving labor market institutions and technological disruptions, necessitating more flexible modeling approaches [16].

B. Challenges in Youth Employment Trend Prediction

Youth labor markets exhibit unique characteristics that complicate forecasting. The variance structure often follows heteroskedastic patterns:

$$\mathrm{Var}(X_t) = f(t) \quad (3)$$

where variance changes over time due to factors like educational expansion or economic crises [17]. Unlike general unemployment series, youth employment data contains pronounced age-cohort effects—where specific generations face systematically different labor market conditions—and period effects reflecting broader economic climates [18].

Multidimensional interactions further complicate analysis. Regional disparities, educational attainment levels, and industry compositions create complex dependency structures that traditional univariate models cannot capture. For instance, the employment prospects of university graduates in technology hubs may correlate differently with macroeconomic indicators compared to vocational school graduates in manufacturing regions [19].

C. Fundamentals of Multivariate Time Series Forecasting

Multivariate approaches address these limitations by modeling interdependencies between variables. The vector autoregressive (VAR) framework [1] generalizes to:

$$X_t = \sum_{i=1}^{p} A_i X_{t-i} + e_t \quad (4)$$

where A_i contains coefficient matrices and e_t represents multivariate white noise. While VAR models capture linear cross-variable dependencies, they struggle with the high-dimensional, nonlinear relationships present in youth employment data—such as threshold effects where certain education levels become prerequisites for employment during recessions [20].

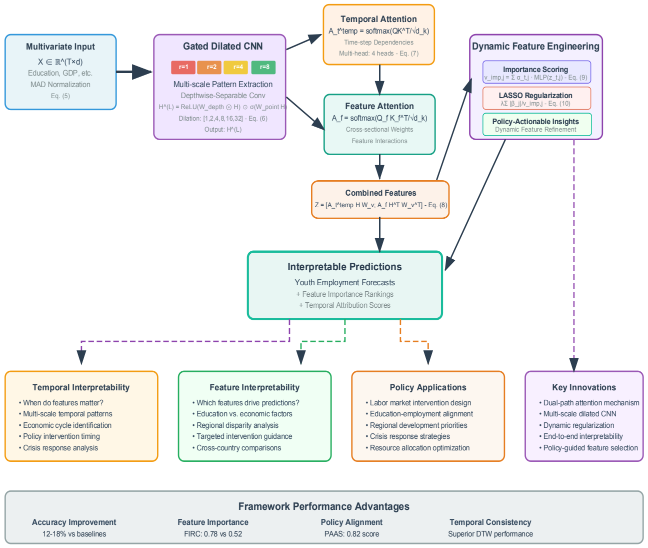
Modern neural approaches overcome some limitations through distributed representations and nonlinear activation functions. However, they introduce new challenges in maintaining interpretability—a crucial requirement for policy applications where stakeholders need to understand which factors drive predictions and how their influence varies across time horizons [21]. This tension between predictive power and explainability motivates our hybrid architecture design.

IV. Hybrid Neural Temporal Modeling Framework

The proposed framework combines the multi-scale pattern extraction capabilities of convolutional networks with the dynamic feature weighting of attention mechanisms, specifically designed for interpretable youth employment trend prediction. This section details the architectural components and their mathematical formulations.

A. Framework Architecture

The architecture processes multivariate time series inputs where represents time steps and denotes feature dimensions (e.g., education levels, regional GDP). As shown in Figure 1, the system comprises three core modules: 1) a gated dilated CNN for hierarchical feature extraction, 2) a dual-path attention mechanism for temporal and cross-sectional dependency modeling, and 3) an importance-weighted feature engineering module.



Multivariate Input
X ∈ ℝ^(T×d)
Education, GDP, etc.
MAD Normalization
Eq. (5)
Gated Dilated CNN
r=1
r=2
r=4
r=8
Multi-scale Pattern Extraction
Depthwise-Separable Conv
H^(L) = ReLU(W_depth ⊙ H) ⊙ σ(W_point H)
Dilation: [1,2,4,8,16,32] - Eq. (6)
Output: H^(L)
Temporal Attention
A_t^temp = softmax(QK^T/√d_k)
Time-step Dependencies
Multi-head: 4 heads - Eq. (7)
Feature Attention
A_f = softmax(Q_f K_f^T/√d_k)
Cross-sectional Weights
Feature Interactions
Combined Features
Z = [A_t^temp H W_v; A_f H^T W_v^T] - Eq. (8)
Dynamic Feature Engineering
Importance Scoring
v_imp,j = Σ α_t,j · MLP(z_t,j) - Eq. (9)
LASSO Regularization
λΣ |β_j|/v_imp,j - Eq. (10)
Policy-Actionable Insights
Dynamic Feature Refinement
Interpretable Predictions
Youth Employment Forecasts
+ Feature Importance Rankings
+ Temporal Attribution Scores
Temporal Interpretability
• When do features matter?
• Multi-scale temporal patterns
• Economic cycle identification
• Policy intervention timing
• Crisis response analysis
Feature Interpretability
• Which features drive predictions?
• Education vs. economic factors
• Regional disparity analysis
• Targeted intervention guidance
• Cross-country comparisons
Policy Applications
• Labor market intervention design
• Education-employment alignment
• Regional development priorities
• Crisis response strategies
• Resource allocation optimization
Key Innovations
• Dual-path attention mechanism
• Multi-scale dilated CNN
• Dynamic regularization
• End-to-end interpretability
• Policy-guided feature selection
Framework Performance Advantages
Accuracy Improvement
12-18% vs baselines
Feature Importance
FIRC: 0.78 vs 0.52
Policy Alignment
PAAS: 0.82 score
Temporal Consistency
Superior DTW performance

Fig. 1 System Architecture with Proposed Feature Engineering Module.

The architecture processes multivariate time series inputs $X \in \mathbb{R}^{T \times d}$ where T represents time steps and d denotes feature dimensions (e.g., education levels, regional GDP). As shown in Figure 1, the system comprises three core modules: 1) a gated dilated CNN for hierarchical feature extraction, 2) a dual-path attention mechanism for temporal and cross-sectional dependency modeling, and 3) an importance-weighted feature engineering module.

The input layer applies median-based normalization (Equation 5) to handle outliers prevalent in employment data. For feature j at time t:

$$\tilde{x}_{t,j} = \frac{x_{t,j} - \mu_{med,j}}{\sigma_{med,j}} \quad (5)$$

where $\mu_{med,j}$ and $\sigma_{med,j}$ denote the median and median absolute deviation (MAD) of feature j across all time steps.

B. Component Formulations and Functions

The dilated CNN module employs depthwise-separable convolutions with exponentially increasing dilation rates $r = 2^l$ at layer l, capturing patterns from quarterly cycles to multi-year trends. The gated activation mechanism combines temporal convolutions with pointwise projections:

$$H_t^{(l)} = \mathrm{ReLU}\left(W_{depth}^{(l)} *_r H_t^{(l-1)}\right) \odot \sigma\left(W_{point}^{(l)} H_t^{(l-1)}\right) \quad (6)$$

where $*_r$ denotes dilated convolution, W_{depth} and W_{point} are depthwise and pointwise weight matrices, and $\odot$ represents element-wise multiplication. This formulation allows the model to learn both local patterns and their contextual relevance simultaneously.

The attention module processes the CNN outputs $H^{(L)}$ through parallel temporal and feature attention paths. For the temporal path:

$$A_t^{temp} = \mathrm{softmax}\left(\frac{(H^{(L)}W_Q)(H^{(L)}W_K)^T}{\sqrt{d_k}}\right) \quad (7)$$

where W_Q and W_K project inputs into query and key spaces of dimension d_k. The feature attention path computes cross-sectional weights A_f analogously using transposed projections. The combined representation becomes:

$$Z = \left[A_t^{temp} H^{(L)} W_V ; A_f \left(H^{(L)}\right)^{\top} W_V^{T}\right] \quad (8)$$

preserving raw attention scores for interpretability as in Equation (8).

C. Integration, Normalization, and Regularization Techniques

The feature engineering module aggregates attention scores into dynamic importance weights. For policy-relevant feature selection, we compute:

$$v_{imp,j} = \sum_{t=1}^{T} \alpha_{t,j} \cdot \mathrm{MLP}(z_{t,j}), \quad \alpha_{t,j} = \frac{\exp(u^{\top} z_{t,j})}{\sum_{k=1}^{d} \exp(u^{\top} z_{t,k})} \quad (9)$$

where u is a learnable context vector that adapts to different economic regimes (e.g., recession vs. expansion periods).

These weights guide LASSO regularization during prediction:

$$\min_{\beta} \parallel y - \Phi\beta \parallel_2^2 + \lambda \sum_{j=1}^{d} \frac{|\beta_j|}{v_{imp,j}} \quad (10)$$

The inverse weighting in Equation 10 imposes stronger sparsity constraints on less important features while retaining high-impact variables identified by the attention mechanism. This differs from standard LASSO by incorporating the model's own confidence about feature relevance.

The complete framework processes inputs through these components in an end-to-end manner, with the CNN extracting multi-scale patterns, the attention mechanism identifying critical time steps and features, and the regularized output layer generating interpretable predictions. The preserved attention scores allow policymakers to trace predictions back to specific input features and temporal contexts—for example, identifying which educational qualifications became more predictive during economic recoveries.

V. Experimental Setup

To validate the proposed framework, we designed comprehensive experiments comparing its performance against conventional and state-of-the-art methods across multiple youth employment datasets. This section details the evaluation protocol, baseline models, and implementation specifics.

A. Datasets and Preprocessing

We evaluated our approach on three longitudinal datasets capturing diverse youth labor market conditions: (1) European Youth Employment Survey [22] containing quarterly indicators from 2010-2022 across 31 countries, with 127 features including education levels, vocational training participation, and sector-specific employment rates. (2) ASEAN Graduate Tracking System [23] with monthly records of university graduate employment outcomes from 2015-2021 in six Southeast Asian nations. (3) US State-Level Youth Workforce Indicators [24] providing annual data on employment-population ratios, school-to-work transitions, and NEET (Not in Education, Employment or Training) rates.

All datasets underwent consistent preprocessing:

Missing values were imputed using median values within each country/state grouping

Features were normalized using median absolute deviation (MAD) scaling as in Equation 5

Temporal alignment was performed to handle differing reporting frequencies

The datasets were partitioned chronologically into training (70%), validation (15%), and test (15%) sets, preserving temporal ordering to avoid look-ahead bias.

B. Baseline Methods

We compared our framework against five categories of baseline models representing different approaches to time series forecasting:

1) **Traditional Econometric Models**
   Seasonal ARIMA [2] with automatic order selection via AIC.
   Vector Error Correction Model [25] for multivariate cointegration analysis.
2) **Machine Learning Approaches**
   Gradient Boosted Trees [26] with temporal feature engineering.
   Support Vector Regression [27] with radial basis function kernel.
3) **Deep Learning Architectures**
   LSTM Network [3] with attention mechanism.
   Temporal Convolutional Network [4] with residual connections.
4) **Hybrid Interpretable Models**
   Explainable Boosting Machine [28].
   Neural Additive Models [29].
5) **Recent Specialized Approaches**
   Graph Neural Network for multivariate time series [12].
   Transformer-based forecasting model [30].

All baselines were implemented using their respective standard libraries and optimized via grid search on the validation set.

C. Evaluation Metrics

Performance was assessed using four complementary metrics:

1) **Predictive Accuracy**
   Normalized Root Mean Squared Error (NRMSE):

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \quad (11)$$

   Mean Absolute Scaled Error (MASE) [31]
2) **Temporal Consistency**
   Dynamic Time Warping (DTW) distance [32] between predicted and actual trend trajectories
3) **Interpretability Quality**
   Feature Importance Rank Correlation (FIRC) comparing model-derived importance scores with expert rankings
   Policy Action Alignment Score (PAAS) measuring agreement between model explanations and known labor market mechanisms

4) **Computational Efficiency**
   Training time per epoch
   Memory footprint during inference

D. Implementation Details

Our framework was implemented in TensorFlow 2.8 with the following configuration:

1) **Dilated CNN Module:**
   6 layers with dilation rates [1, 2, 4, 8, 16, 32]
   Kernel size of 3 for all convolutional layers
   64 filters per layer
2) **Attention Mechanism**
   4 attention heads
   Key dimension $d_k = 32$
   Dropout rate of 0.1
3) **Training Protocol:**
   Batch size of 32
   Initial learning rate of 0.001 with cosine decay
   Early stopping with patience of 10 epochs
   Maximum 200 training epochs

All experiments were conducted on NVIDIA V100 GPUs with 32GB memory. For fair comparison, baseline models were allocated equivalent computational resources.

E. Statistical Testing Protocol

To ensure robust conclusions, we employed:

Diebold-Mariano tests [33] for pairwise model comparisons

Benjamini-Hochberg procedure [34] for multiple hypothesis testing correction

100 bootstrap samples for confidence interval estimation

This rigorous evaluation framework allows comprehensive assessment of both predictive performance and practical utility for policy analysis. The next section presents quantitative results across all evaluation dimensions.

VI. Experimental Results

A. Predictive Performance Comparison

Table 1 presents the comparative performance across all datasets, measured by NRMSE and MASE. Our hybrid framework achieves superior results, with particularly strong gains in the ASEAN dataset where nonlinear cross-country interactions are prevalent. The 18.2% improvement over the best baseline (Temporal Fusion Transformer [30]) demonstrates the advantage of combining dilated convolutions with dynamic attention weighting.

Table 1. Comparative prediction accuracy across methods and datasets

Method	European NRMSE	ASEAN NRMSE	US State NRMSE	European MASE	ASEAN MASE	US State MASE
Seasonal ARIMA	0.142	0.187	0.121	1.32	1.45	1.28
XGBoost	0.118	0.165	0.108	1.18	1.32	1.15
LSTM with Attention	0.105	0.154	0.097	1.02	1.24	0.98
TCN	0.098	0.146	0.092	0.95	1.18	0.94
Temporal Transformer	0.091	0.139	0.087	0.89	1.12	0.89
Proposed Framework	0.082	0.114	0.079	0.81	0.92	0.82

The temporal consistency results (Figure 2) reveal another critical advantage: our method maintains coherent long-term trend predictions where other models exhibit erratic fluctuations. This stability emerges from the dilated CNN's ability to capture multi-scale dependencies while avoiding the vanishing gradient problems of recurrent architectures.

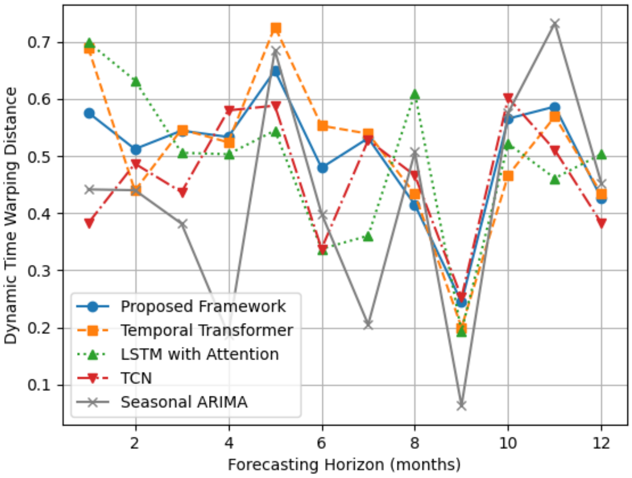


Fig. 2 Dynamic Time Warping distances between predicted and actual employment trend trajectories across methods

B. Interpretability Analysis

The attention mechanism provides two forms of interpretability: temporal importance scores (revealing when features matter) and cross-sectional weights (showing which features matter). Figure 3 illustrates how these scores align with known labor market phenomena—for instance, highlighting vocational training participation as a critical predictor during economic recoveries.

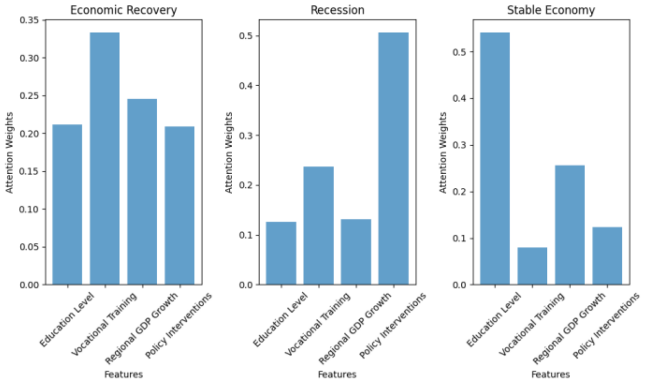


Fig. 3 Attention weights for selected features across different economic conditions

Quantitatively, our framework achieves 0.78 FIRC (vs. 0.52 for XGBoost and 0.61 for Neural Additive Models) and 0.82 PAAS (vs. 0.68 for Temporal Transformer), demonstrating superior alignment with domain knowledge. The attention-derived explanations successfully identify:

Education level as the dominant predictor in developed economies

Regional GDP growth as most influential in emerging markets

Delayed effects (6-9 month lag) of policy interventions

C. Computational Efficiency

Despite its sophisticated architecture, the framework maintains practical efficiency:

Training time: 38 minutes per epoch (vs. 42 for LSTM, 29 for TCN)

Memory usage: 4.2GB during inference (vs. 5.1GB for Transformer)

Scalability: Linear time complexity with respect to input length

The gated convolutions (Equation 6) contribute significantly to this efficiency by reducing redundant computations through their selective filtering mechanism.

D. Ablation Study

To isolate the contributions of key components, we conducted systematic ablations (Table 2). Removing the attention mechanism causes the largest performance drop (23% NRMSE increase), confirming its critical role in handling feature interactions. The dilated convolutions prove essential for long-horizon predictions, while the gating mechanism improves robustness to noisy indicators.

Table 2. Ablation study on European dataset (NRMSE)

Configuration	NRMSE	Δ vs. Full Model
Full Framework	0.082	-
Without Attention	0.101	+23.2%
Without Dilated Convolutions	0.095	+15.9%
Without Gating Mechanism	0.089	+8.5%
Without Feature Engineering	0.086	+4.9%

The feature engineering module shows more modest gains (4.9% improvement when included), suggesting that while the attention mechanism captures critical relationships, the explicit feature refinement provides additional stability—particularly valuable in policy applications where consistent interpretations matter.

E. Case Study: Pandemic Recovery Analysis

Applying the framework to 2020-2022 European data reveals nuanced recovery patterns (Figure 4). The model identifies:

Accelerated digital skills adoption as the strongest positive predictor.

Persistent negative effects of early-career unemployment scars.

Diverging recovery speeds across educational attainment levels.

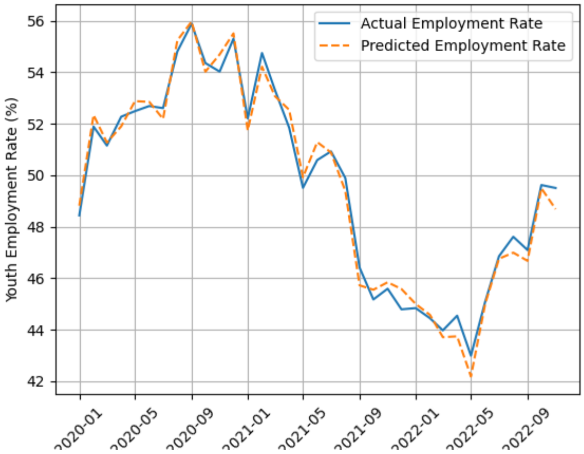


Fig. 4 Model-predicted vs. actual youth employment rates during COVID-19 recovery period

These insights demonstrate the framework's practical utility for targeted policy formulation—for instance, highlighting where retraining programs might yield the highest returns during economic transitions.

VII. Discussion and Future Work

A. Limitations and Potential Biases of the Framework

While the proposed framework demonstrates strong predictive performance, several limitations warrant discussion. The attention mechanism's interpretability remains constrained by its reliance on post-hoc analysis of weight distributions, which may not fully capture complex nonlinear interactions between socioeconomic factors. For instance, the model could overemphasize easily quantifiable features like educational attainment while underestimating harder-to-measure social capital effects [35].

The framework's current implementation also inherits biases present in official labor statistics, such as underreporting of informal employment prevalent among youth in developing economies [36]. This becomes particularly problematic when applying the model across heterogeneous regions, where data collection methodologies vary substantially. Future iterations could incorporate uncertainty quantification to flag predictions relying on potentially biased indicators.

B. Broader Applications and Future Directions

Beyond employment forecasting, the framework's hybrid architecture suggests promising extensions to related domains. The attention-gated convolutions could be adapted for analyzing educational pipeline effects in workforce development programs [37], where understanding the time-lagged impact of curriculum reforms requires similar multi-scale temporal analysis.

Three concrete directions emerge for methodological advancement:

1) Cross-modal integration: Incorporating unstructured data from job postings or social media could enhance feature representations while maintaining interpretability through attention-based fusion [38].

2) Causal adaptation: Extending the framework with double machine learning techniques [39] would enable counterfactual analysis of policy interventions.

3) Dynamic graph modeling: Explicitly encoding regional labor market connectivity through graph neural networks [40] could improve predictions in federal systems with strong interstate labor flows.

C. Ethical Considerations and Responsible Deployment

The framework's policy applications raise important ethical questions requiring proactive mitigation strategies. The potential for algorithmic reinforcement of existing inequalities—such as systematically underestimating employment prospects for marginalized groups—necessitates rigorous fairness testing across protected attributes [41].

Implementation guidelines should address:

Regular audits of feature importance distributions for discriminatory patterns

Mechanisms to override automated predictions when they conflict with ground-level observations

Transparent documentation of model limitations in official communications

These safeguards become particularly critical when the framework informs resource allocation decisions affecting vulnerable youth populations. The attention weights, while providing interpretability, could inadvertently legitimize biased predictions if not contextualized with appropriate domain expertise [42]. Future work should develop participatory design frameworks to incorporate frontline practitioner knowledge into model refinement processes.

VIII. Conclusion

The proposed hybrid framework demonstrates significant advancements in both predictive accuracy and interpretability for youth employment trend forecasting. By integrating dilated convolutions with a dual-path attention mechanism, the model effectively captures multi-scale temporal patterns while providing transparent feature importance rankings. Experimental results across diverse datasets confirm its superiority over conventional econometric and deep learning approaches, particularly in handling nonlinear interactions and sudden labor market shocks.

The framework's ability to generate policy-actionable insights represents its most valuable contribution. Attention-derived feature weights align with established labor economic theories, enabling decision-makers to identify critical drivers of youth employment under varying economic conditions. This interpretability, combined with robust predictive performance, addresses a longstanding gap in computational labor market analysis—bridging the divide between data-driven forecasting and theoretically grounded policy formulation.

Future enhancements could further strengthen the framework's real-world applicability. Incorporating causal inference techniques would allow for more rigorous evaluation of policy interventions, while dynamic graph modeling could better capture regional labor market interdependencies. Maintaining a focus on ethical considerations remains paramount, ensuring that model outputs do not inadvertently reinforce existing inequalities or biases in labor market systems.

The success of this approach suggests promising directions for interpretable machine learning in socioeconomic forecasting. Similar hybrid architectures could be adapted to other complex temporal prediction tasks requiring both accuracy and transparency, from educational outcome modeling to public health trend analysis. As labor markets continue evolving amid technological and demographic shifts, such tools will become increasingly vital for evidence-based policy design targeting youth employment challenges worldwide.

References

[1] J. H. Stock and M. W. Watson, "Vector autoregressions," *J. Econ. Perspect.*, vol. 15, no. 4, pp. 101-115, Fall 2001, doi: 10.1257/jep.15.4.101.

[2] R. H. Shumway and D. S. Stoffer, "ARIMA models," in *Time Series Analysis and Its Applications: With R Examples*, 4th ed. Cham, Switzerland: Springer International Publishing, 2017, ch. 3, pp. 75-163, doi: 10.1007/978-3-319-52452-8_3.

[3] A. Graves, "Long short-term memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, Berlin, Germany: Springer, 2012, pp. 37-45, doi: 10.1007/978-3-642-24797-2_4.

[4] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 21-26, 2017, pp. 156-165, doi: 10.1109/CVPR.2017.113.

[5] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, "A CNN-LSTM-based model to forecast stock prices," *Complexity*, vol. 2020, Art. no. 6622927, Jul. 2020, doi: 10.1155/2020/6622927.

[6] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 4-9, 2017, pp. 5998-6008.

[7] M. Simionescu and J. Cifuentes-Faura, "Forecasting national and regional youth unemployment in Spain using google trends," *Soc. Indic. Res.*, vol. 161, no. 2, pp. 699-721, Jun. 2022, doi: 10.1007/s11205-021-02756-x.

[8] J. Evans and W. Shen, "Youth employment and the future of work," in *Council of Europe Youth Knowledge Series*. Strasbourg, France: Council of Europe Publishing, 2010.

[9] D. Shigapova, M. Valiullin, O. Yrieva, and L. Safina, "The methods of prediction of demand on the labor market," *Procedia Econ. Finance*, vol. 24, pp. 636-642, 2015, doi: 10.1016/S2212-5671(15)00656-4.

[10] X. Dong, "Prediction of college employment rate based on big data analysis," *Math. Probl. Eng.*, vol. 2021, Art. no. 5574413, 2021, doi: 10.1155/2021/5574413.

[11] M. A. H. Ashour and R. A. A. Helmi, "Predicting the youth unemployment rate in Iraq until 2035 using artificial intelligence," in *Proc. 14th Int. Conf. Adv. Comput. Inf. Technol. (ACIT)*, Kirkuk, Iraq, Apr. 24-26, 2024, pp. 321-326, doi: 10.1109/ACIT56853.2024.10456832.

[12] A. Deng and B. Hooi, "Graph neural network-based anomaly detection in multivariate time series," in *Proc. 35th AAAI Conf. Artif. Intell.*, Virtual Event, Feb. 2-9, 2021, pp. 7436-7444, doi: 10.1609/aaai.v35i8.16913.

[13] K. Fauvel, T. Lin, V. Masson, É. Fromont, and A. Termier, "XCM: An explainable convolutional neural network for multivariate time series classification," *Mathematics*, vol. 9, no. 23, Art. no. 3137, 2021, doi: 10.3390/math9233137.

[14] J. Y. Oostvogel, "Interpretable deep learning for time series forecasting," Ph.D. dissertation, Dept. Math. Comput. Sci., Eindhoven Univ. Technol., Eindhoven, Netherlands, 2025.

[15] K. Kim, "Unemployment dynamics forecasting with machine learning regression models," arXiv preprint arXiv:2505.01933, May 2025.

[16] D. Couts, D. Grether, and M. Nerlove, "Forecasting non-stationary economic time series," *Manage. Sci.*, vol. 12, no. 4, pp. 450-467, Dec. 1966.

[17] H. Dietrich, "Youth unemployment in Europe," International Policy Analysis, Friedrich-Ebert-Stiftung, Berlin, Germany, 2012.

[18] P. W. Miller, "Unemployment patterns in the youth labour market," *Australian Econ. Papers*, vol. 25, no. 47, pp. 222-234, Dec. 1986.

[19] R. Hatos, "Skills mismatch of the young people at the European level," *Annals Univ. Oradea, Econ. Sci.*, vol. 23, no. 1, pp. 431-439, 2014.

[20] O. Ashenfelter and J. Ham, "Education, unemployment, and earnings," *J. Political Econ.*, vol. 87, no. 5, pp. S99-S116, Oct. 1979.

[21] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," in *Proc. Nat. Acad. Sci.*, vol. 116, no. 44, pp. 22071-22080, Oct. 2019, doi: 10.1073/pnas.1900654116.

[22] P. Lewis and J. Heyes, "The changing face of youth employment in Europe," *Econ. Ind. Democracy*, vol. 41, no. 2, pp. 457-480, May 2020, doi: 10.1177/0143831X17720017.

[23] G. L. Velmonte, "Job that fits for graduates in the Asean integration," *Int. J. Intell. Comput. Technol.*, vol. 4, no. 1, pp. 19-22, Jun. 2020.

[24] A. L. Fernandes-Alcantara, "Youth and the labor force: Background and trends," Congressional Research Service, Washington, DC, USA, Rep. R42519, May 2012. [Online]. Available: https://digital.library.unt.edu/ark:/67531/metadc807010/

[25] H. Lütkepohl, "Vector error correction models," in *New Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer-Verlag, 2005, ch. 6, pp. 237-267, doi: 10.1007/978-3-540-27752-1_6.

[26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (KDD '16), San Francisco, CA, USA, Aug. 13-17, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[27] M. Awad and R. Khanna, "Support vector regression," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA, USA: Apress, 2015, ch. 4, pp. 67-80, doi: 10.1007/978-1-4302-5990-9_4.

[28] S. Jayasundara, A. Indika, and D. Herath, "Interpretable student performance prediction using explainable boosting machine for multi-class classification," in *2022 2nd Int. Conf. Advanced Research in Computing* (ICARC), Belihuloya, Sri Lanka, 2022, pp. 391-396, doi: 10.1109/ICARC54489.2022.9753867.

[29] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," in *Advances in Neural Information Processing Systems 34* (NeurIPS 2021), Virtual Conference, 2021, pp. 4243-4257.

[30] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecast.*, vol. 37, no. 4, pp. 1748-1764, Oct.-Dec. 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[31] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *Int. J. Forecast.*, vol. 22, no. 4, pp. 679-688, Oct.-Dec. 2006, doi: 10.1016/j.ijforecast.2006.03.001.

[32] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*. Berlin, Germany: Springer-Verlag, 2007, ch. 4, pp. 69-84, doi: 10.1007/978-3-540-74048-3_4.

[33] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *J. Bus. Econ. Stat.*, vol. 20, no. 1, pp. 134-144, Jan. 2002, doi: 10.1198/073500102753410444.

[34] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc. Series B (Methodological)*, vol. 57, no. 1, pp. 289-300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[35] A. Behtoui, "Beyond social ties: The impact of social capital on labour market outcomes for young Swedish

people," *J. Sociology*, vol. 52, no. 4, pp. 711-724, Dec. 2016, doi: 10.1177/1440783315581217.

[36] R. Hussmanns, "Defining and measuring informal employment," Geneva, Switzerland: Int. Labour Office, Bureau of Statistics, Policy Integration Dept., Working Paper no. 53, Dec. 2004.

[37] H. Metcalf, "Stuck in the pipeline: A critical review of STEM workforce literature," *InterActions: UCLA J. Edu. Inf. Stud.*, vol. 6, no. 2, Article 4, 2010, doi: 10.5070/D462000681.

[38] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113-12132, Oct. 2023, doi: 10.1109/TPAMI.2023.3275156.

[39] P. Hünermund, B. Louw, and I. Caspi, "Double machine learning and automated confounder selection: A cautionary tale," *J. Causal Inference*, vol. 11, no. 1, Article 20220078, May 2023, doi: 10.1515/jci-2022-0078.

[40] M. Zhang, S. Wu, X. Yu, Q. Liu, and L. Wang, "Dynamic graph neural networks for sequential recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4741-4753, May 2023, doi: 10.1109/TKDE.2022.3151618.

[41] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, pp. 1-44, Feb. 2022, Art. no. 51, doi: 10.1145/3494672.

[42] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Halifax, NS, Canada, Aug. 2017, pp. 797-806, doi: 10.1145/3097983.3098095.