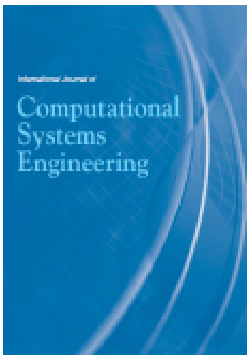
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Computational Systems Engineering

ISSN online: 2046-3405 - ISSN print: 2046-3391
https://www.inderscience.com/ijcsyse

Causal transformer and counterfactual reasoning: deconstruction analysis of the impact of teaching strategies on academic achievement

Lipeng Zhang

DOI: 10.1504/IJCSYSE.2025.10075320

Article History:

Received:	12 August 2025
Last revised:	12 October 2025
Accepted:	23 October 2025
Published online:	07 January 2026

Causal transformer and counterfactual reasoning: deconstruction analysis of the impact of teaching strategies on academic achievement

Lipeng Zhang

School of Marxism,
Shijiazhuang College of Applied Technology,
Shijiazhuang City, Hebei Province, 050800, China
Email: lipengzhang0606@163.com

Abstract: Modern education increasingly relies on data-driven decision-making, requiring causal inference methods to assess teaching strategies beyond correlations. Challenges such as time-varying confounders, unobserved counterfactuals, temporal dependencies of interventions, and heterogeneous responses limit strategy design and evaluation. Existing methods also struggle with temporal dynamics and complex causal structures. To address these issues, causal temporal contextual reasoning (CTCR) was developed, incorporating a dynamic disentanglement mechanism for time-varying confounders, a two-way causal representation module, and a counterfactual generation algorithm constrained by temporal logic. Experiments on higher education (Dataset-H), K12, and MOOCs datasets show CTCR's effectiveness. On Dataset-H, it achieves MSE-T 5.53 × 10^{-2}, PEHE 5.16 × 10^{-1}, and CP@K 0.89, outperforming comparative models. Performance volatility across K12 and MOOCs is ≤ 29.3%, and CTCR remains robust under 30% data sparsity and 5 dB noise, demonstrating strong generalisation and reliability.

Keywords: causal transformer; counterfactual reasoning; instructional strategy; CTCR; academic achievement; time-varying confounders.

Reference to this paper should be made as follows: Zhang, L. (2025) 'Causal transformer and counterfactual reasoning: deconstruction analysis of the impact of teaching strategies on academic achievement', *Int. J. Computational Systems Engineering*, Vol. 9, No. 17, pp.1–24.

Biographical notes: Lipeng Zhang is working in Hijiazhuang College of Applied Technology. He graduated from Guizhou Minzu University in 2012 with criminal law major. Currently, he works in Shijiazhuang College of Applied Technology as a Lecturer in the School of Marxism. His main research interests are higher education, vocational education, etc.

1 Introduction

In modern education, data – driven decision – making necessitates causal inference that transcends mere correlation to precisely evaluate the genuine impact of teaching strategies on academic performance (Kitto et al., 2023). Nevertheless, the majority of current mainstream methods rely on the statistical associations within observational data, rendering it arduous to disentangle the true causal effects from the interference of confounding factors. The reasons are as follows: the coupling effect of time – varying confounders, such as the dynamic alteration of students' autonomous learning ability, which not only influences the selection of teaching strategies but also correlates with academic performance (Marantika, 2021); counterfactual outcomes are unobservable, and traditional methods are unable to address the query of ‘how would students' academic performance vary if an alternative strategy were employed’ (Keller and Branson, 2024); the temporal dependence and heterogeneous responses of educational interventions augment the complexity of identification.

To address these challenges, educational causal research has undergone iterative methodological evolution. Early approaches relied on quasi-experimental designs and parametric models, such as difference-in-differences (DID) (Tournaki, 2023), instrumental variables (IV) (Alauddin et al., 2017), and matching methods (Keele et al., 2021). While these established the research foundation, they overlooked dynamic confounding, feedback loops, and temporal logic constraints of counterfactual paths, failing to support the deconstruction of complex instructional strategy effects.

In recent years, the rapid advancement of artificial intelligence has enabled scholars to integrate deep learning with causal inference, offering new approaches for handling temporal dependencies and nonlinear relationships. Causal forests can identify heterogeneous treatment effects but inadequately capture temporal dependencies (Chen et al., 2021). Bayesian neural networks (BNNs) suit small-sample scenarios but face computational bottlenecks with high-dimensional data (Park et al., 2021). Additionally, transformer-based models like CausalBERT (Li et al., 2021) and G-transformer (Xiong et al., 2024) partially mitigate interference from time-varying confounders in causal effect estimation.

Recent progress focuses on dynamic modelling and instrumental variable innovations. Causal transformer's segmented attention mechanism decomposes long sequences into short segments, reducing computational load while capturing dynamic intervention effects. However, its reliance on preset time granularity struggles with non-uniform temporal variations (Zhu et al., 2024). MATTE incorporates domain knowledge to constrain counterfactual generation, reducing bias from missing data. Yet its dependence on expert-defined rules limits cross-domain transferability (Yan et al., 2023). CB-IV (Wu et al., 2022) and AutoIV (Yuan et al., 2022) automate instrumental variable selection, alleviating selection bias in high-dimensional settings. However, theoretical validation of their efficacy remains pending.

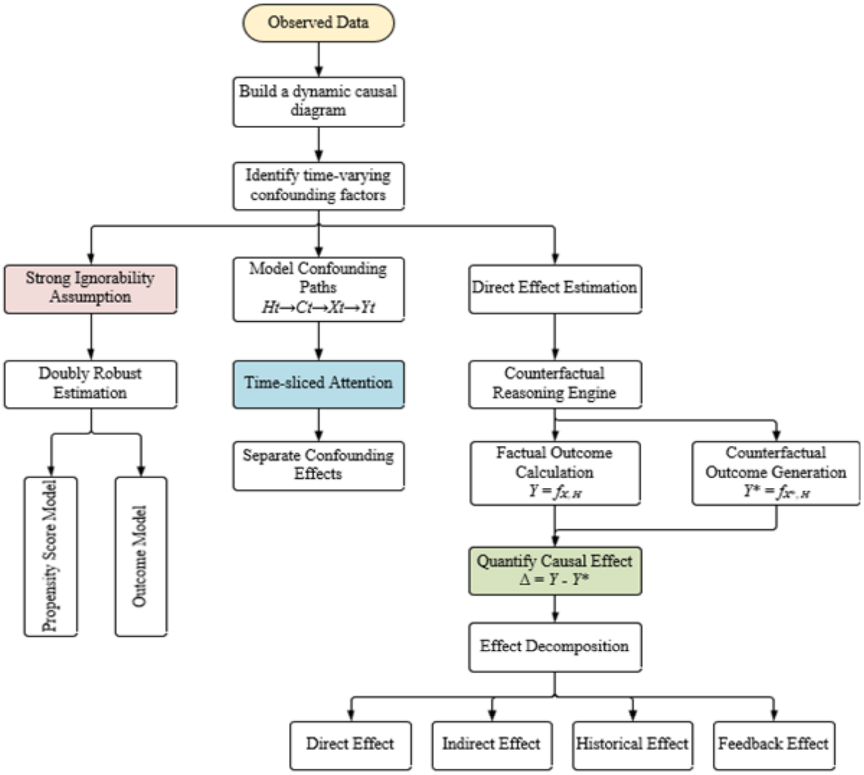
In essence, existing analytical methods encounter difficulties in satisfying the profound requirements of educational causal inference, and they face the following challenges: time – varying confounders contravene the strong ignorability assumption, leading traditional static models to yield biases because of the disregard for the time dimension. The long – range dependence and feedback loops of intervention effects render the analysis of existing architectures, which assume no feedback, ineffective. The temporal logic constraints of counterfactual reasoning are in conflict with the traditional consistency assumption, causing the counterfactual paths often generated to run counter to teaching logic. Consequently, this study develops CTCR to accomplish a meticulous deconstruction of the effects of teaching strategies on academic performance. The main contributions are as follows:

1 A dynamic confounder disentanglement mechanism using Transformer multi-head attention to isolate confounders from direct intervention effects, resolving causal confounding induced by overlooked temporal dependencies

2 A bidirectional causal representation module concurrently modelling feedforward impacts and feedback regulation of instructional strategies, overcoming unidirectional causal assumptions in static models

3 A temporally-constrained counterfactual generation algorithm embedding educational temporal dependency rules into potential outcome models, ensuring counterfactual paths align with educational practice logic.

Figure 1 The process of counterfactual reasoning (see online version for colours)



2 Theoretical foundations

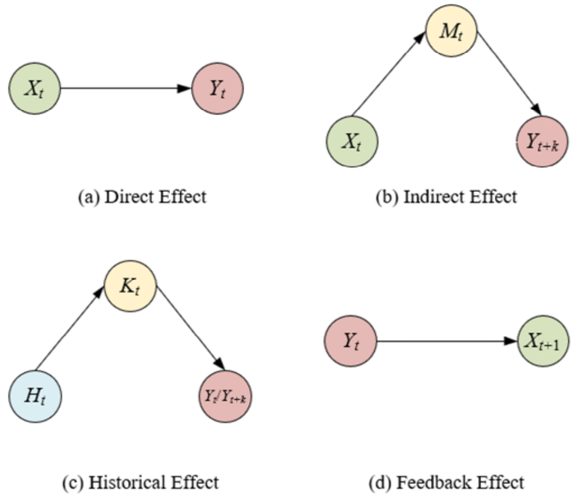
2.1 Counterfactual reasoning and causal inference

Counterfactual reasoning serves as a core methodology within causal inference, with the objective of exploring the question: ‘what would the outcome be if the intervention had not occurred?’ (Nyhout and Ganea, 2021). In the realm of educational intervention evaluation, its significance lies in surmounting the innate limitations of observational data. Specifically, this pertains to the incapacity to directly observe the potential outcomes of the same individual under diverse strategies (Sinha and Kapur, 2021).

More precisely, for a group of students who have undergone a particular teaching strategy *T*, their actual academic performance *Y* represents the ‘factual’ outcome that is observed. Conversely, counterfactual reasoning endeavors to estimate the potential academic performance Y^* of this group had they received an alternative strategy T' or no intervention at all. This ‘contrast difference’ ($Y–Y^*$) constitutes the theoretical foundation for assessing the causal effect of the intervention.

Nonetheless, the time –varying confounders frequently encountered in educational scenarios contravene the strong ignorability assumption of traditional causal inference. As a result, this leads to substantial biases in counterfactual estimates derived from static observations (Hong and Raudenbush, 2008). Figure 1 illustrates the process of counterfactual reasoning.

Figure 2 Causal path (a) direct effect (b) indirect effect (c) historical effect (d) feedback effect (see online version for colours)



(a) Direct Effect

(b) Indirect Effect

(c) Historical Effect

(d) Feedback Effect

2.2 Time series effect decomposition

Figures 2(a)–2(d) illustrate causal pathways at distinct time slices *t*. The total effect (TE) of instructional strategies on academic performance decomposes into four components.

- Direct effect (DE): immediate impact of current strategy X_t on contemporaneous performance Y_t, following path $X_t \rightarrow Y_t$.
- Indirect effect (IE): effect of current strategy X_t on subsequent performance $Y_t + k$ mediated through intermediate variable M_t, via path $X_t \rightarrow M_t \rightarrow Y_t + k$.
- Historical effect (HE): persistent influence of historical strategy sequence H_t on current/future performance by shaping current state K_t, through path $H_t \rightarrow K_t \rightarrow Y_t/Y_t + k$.

- Feedback effect (FE): closed-loop path where current outcome Y_t inversely affects future intervention decisions $X_t + 1$, denoted $Y_t \rightarrow X_t + 1$.

2.3 Temporal counterfactual definition

Under the dynamic causal graph framework, counterfactual reasoning must strictly adhere to temporal logic constraints of interventions (Barbero et al., 2023). We define:

- Factual outcome: observed academic performance $\underline{Y_{h,x}}$ when implementing instructional strategy $X_t = x$.
- Counterfactual outcome: potential performance $Y_{h,x}{}^*$ obtained by replacing current intervention with $X_t = x^*$ ($x^* \neq x$), while holding historical sequence $H_t = h$ constant.
- The conditional treatment effect (CTE) of strategy x relative to x^* is then defined as:

$$CTE\left(x \rightarrow x^* \middle| h\right) = Y_h^{fact}(x) - Y_h^{cf}\left(x^*\right) \tag{1}$$

3 Model design and comparison

3.1 Model design

Figure 3 illustrates the architecture of CTCR, designed with the dual objectives of: disentangling temporal confounding effects of instructional strategies; generating counterfactual paths consistent with educational logic.

- CTCR comprises two phases: clue extraction and hypothesis verification.

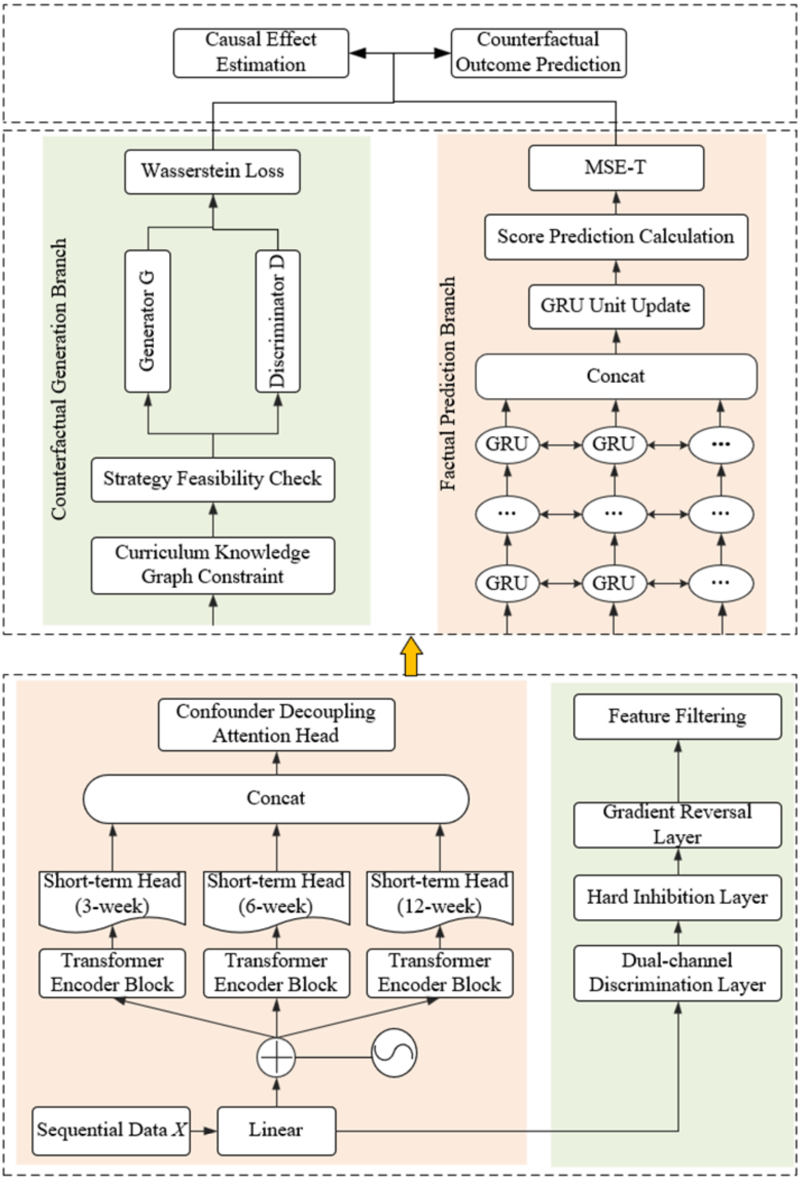
  The clue extraction phase includes a temporal attention network and a spurious correlation filtering block. The temporal attention network inputs pre-processed longitudinal data – instructional strategies, student behaviours, and control variables, segmented weekly into $X = x_1, x_2, \ldots, x_T$. It employs four independent attention heads to capture influences at different temporal scales.

- Short-term head: uses a 3-week sliding window to focus on immediate effects of localised strategies, calculating attention weights between the current week t and the preceding two weeks (t–1, t–2).

$$A_{ij}^{short} = Softmax\left(\frac{Q_i W_Q^{short} \cdot \left(K_j W_K^{short}\right)^{\top}}{\sqrt{d}}\right), j \in [i-2, i] \tag{2}$$

- Here, Q denotes the query vector, representing features of the current week i. K denotes the key vector, representing features of historical week j. W_Q^{short}, W_K^{short} are learnable parameter matrices projecting input features into query and key spaces. $\sqrt{d}$ is a scaling factor controlling the dot product magnitude to prevent gradient vanishing.
- Mid-term head: employs a 6-week sliding window to capture cumulative effects during instructional phases, with attention spanning weeks $[i–5, j]$.

Figure 3 CTCR architecture (see online version for colours)


Causal Effect Estimation
Counterfactual Outcome Prediction
Counterfactual Generation Branch
Wasserstein Loss
Generator G
Discriminator D
Strategy Feasibility Check
Curriculum Knowledge Graph Constraint
Factual Prediction Branch
MSE-T
Score Prediction Calculation
GRU Unit Update
Concat
GRU
GRU
···
···
···
···
GRU
GRU
···
Confounder Decoupling Attention Head
Concat
Short-term Head (3-week)
Short-term Head (6-week)
Short-term Head (12-week)
Transformer Encoder Block
Transformer Encoder Block
Transformer Encoder Block
Sequential Data X
Linear
Feature Filtering
Gradient Reversal Layer
Hard Inhibition Layer
Dual-channel Discrimination Layer

- Long-term head: utilises a 12-week sliding window to analyse global impacts of semester-level strategy sequences, extending attention to weeks $[i{-}11, j]$.
- Confounder disentanglement head: separates interference from time-varying confounders by maximising mutual information $I(T_i; Z_i)$ between interventions and latent representations, while minimising $I(Z_i; X_i)$ to ensure exclusive encoding of confounding factors.

- Spurious correlation filtering module adopts a three-tiered structure: dual-channel discrimination, hard suppression, and gradient reversal. The dual-channel discrimination layer first computes mutual information $I(v, T)$ and correlation coefficients $\rho(v, Y)$ in parallel between variables and instructional interventions/academic performance. Non-causal associations are identified using predetermined thresholds ($I(v, T) < 0.1$ and $| \rho(v, Y) | < 0.15$).
- Based on discrimination results, the system executes two-level processing: for superficially correlated variables meeting suppression conditions, their attention weights are forced to zero in the weight matrix; for control variables, their embedding vectors are directed through a gradient reversal layer with negative scaling ($\lambda = -0.5$) to inversely scale gradients, preventing the model from exploiting them for intervention prediction.
- Hypothesis verification phase contains two branches: factual prediction and counterfactual generation. The factual prediction branch processes real instructional strategy sequences using a gated recurrent temporal predictor. Its core is a bidirectional GRU network: the forward GRU encodes cumulative effects of historical interventions and behaviours, while the backward GRU captures teachers' feedback adjustments based on periodic grades. The GRU unit updates hidden states weekly $h_t = f_{GRU}(h_{t-1,} x_t, T_t)$, outputting estimated grades for week $\hat{Y}^F_{t+1}$. The loss function employs temporally-weighted mean squared error (MSE-T):

$$L_{fact} = \sum_{t=1}^{T-1} a^{T-t} \cdot \left(\hat{Y}^F_{t+1} - Y_{T+1}\right)^2 \tag{3}$$

- Here, $a = 0.9$, give higher weight to recent predictions.
- Counterfactual generation branch constructs virtual intervention paths through a strategy replacement engine and adversarial training. First, hard constraints are enforced based on the curriculum knowledge graph: if a student has not mastered knowledge point A, higher-order strategies dependent on A are prohibited, ensuring counterfactual paths adhere to pedagogical progression logic. Subsequently, a conditional GAN generates counterfactual outcomes:
- Generator G: receives real sequences and noise vectors, outputs strategy-replaced sequence T and predicted outcome $\hat{Y}^{CF}$.
- Discriminator D: constrains the distribution of $\hat{Y}^{CF}$ to approximate real grade distributions via Wasserstein distance.

The model initially allows only single-week strategy replacement, progressively enabling multi-week recombination to achieve dynamic equilibrium, ultimately generating globally consistent counterfactual sequences under graph constraints.

3.2 Comparison of architectural differences

To determine the originality and differences of the CTCR model, five mainstream educational causal models were selected and their architectures were compared.

The architectural design of CTCR is centred around 'the temporal dynamics of educational interventions' and 'the interpretability of causal effects', which is essentially different from existing models in terms of module functions and interaction logics. The framework differences are shown in Table 1.

Table 1 Comparison of architectural differences among different models

Model	*Module composition*	*Temporal processing method*	*Confounder processing module*	*Counterfactual generation logic*	*Feedback mechanism (y→x)*
Causal Forest	Decision tree + hetero nodes	None (temp. indep.)	No module (sample match)	None	None
CFRNet (Deng et al., 2024)	Uni-GRU + CF pred branch	Forward recursive	Static ctrl var. embed	State replace (no edu. cons.)	None
Causal transformer	Temp slice attn + FC layer	Fixed-slice dep.	Static slice separation	Intra-slice strat. replace	None
Multimodal causal transformer (Zhang et al., 2023)	Multimodal embed + causal attn	Multimodal parallel align	Feature fusion	Modality cons. (no edu. rules)	None
GPT-4 Edu. Spec.	Transformer decoder + edu. fine-tune	Linguistic fluency dep.	No module (data distr.)	Text fluency (no prac. cons.)	None
CTCR	Temp attn (4 heads) + bi-GRU + knowledge graph	Hier. dyn window (3/6/12w)	Dyn confounder disentanglement	Knowledge graph + prog. gen (teach. logic)	Yes (backward GRU)

4 Experiments

4.1 Data preparation

To conduct a systematic analysis of the impact of teaching strategies on students' behaviours and academic performance, a multi – dimensional indicator system was developed grounded in the theories of data mining and causal inference. Key variables were categorically classified in a structured manner and operationally defined. Table 2 defines four teaching – intervention variables by means of multi – hot encoding. Table 3 presents the time – series data indicators of students' behaviours, encompassing learning engagement, assignment patterns, interaction quality, and cognitive – behaviour markers. Their measurement methods and educational implications are derived from the generative learning – behaviour analysis framework and metacognitive theory (Dai et al., 2025; Ozturk, 2017; Dennis and Somerville, 2023). Table 4 shows the set of control variables, which includes three types of confounding factors: academic background, learning environment, and social influence. The screening principles adhere to the principles of educational causal inference (Valbuena et al., 2021; Forney and Mueller, 2022), and optimisation was achieved through statistical tests (VIF < 5) and feature – importance assessment (Top 80%).

Table 2 Instructional intervention variables

Variable code	*Instructional strategy*	*Definition*	*Data marker/encoding conditions*
T1	Traditional face-to-face	Teacher-led lectures without online learning tasks	Classroom attendance rate > 90% and no recorded lecture access
T2	Blended learning	Recorded lectures (50%) + in-depth classroom discussions (50%)	Recorded video viewing duration ≥ 30% of total course time and ≥ 2 discussion records/week
T3	Project-based learning (PBL)	Cross-week group tasks (experimental design, research reports, etc.)	≥ 1 group task submission/stage and ≥ 5 collaborative forum discussions/week
T4	Adaptive learning	AI dynamically adjusts content difficulty (re-push of incorrect problems, micro-lectures)	Click-through rate on recommended content > 60% and ≥ 3 learning path jumps/week

Table 3 Temporal student behaviour data

Metric category	*Specific metric*	*Measurement method*
Learning engagement	Video learning effectiveness	Percentage of viewing segments >5 minutes per session (filtering invalid clicks)
Assignment behaviour	Submission behaviour	Negative logarithmic transformation of submission time relative to deadline
Interaction quality	Forum interaction depth	Semantic complexity of questions/answers in course forums
Cognitive markers	Error redo interval	Time difference (days) between first and last submission of wrong answers on same knowledge point

4.2 Data pre-processing and dataset partition

To validate the generalisation capacity of CTCR, three distinct types of datasets, encompassing higher – education institutions, K12 (Martin et al., 2023), and MOOCs (Ani and Khor, 2024), were developed. The comparability of cross – scenario indicators was guaranteed via ‘variable logic alignment’. Table 6 depicts the basic information of these three datasets. Each dataset was divided into training, validation, and test sets using time – series data. Specifically, the training set constituted 70% and was composed of early complete data. The validation set and the test set each made up 15% and were selected from subsequent time – series data. During the partitioning process, the proportion of student groups and the distribution of teaching strategies in each subset were strictly preserved, with a deviation of less than 5% from the original dataset, and a unified pre – processing procedure was implemented.

Table 4 Control variable set

Category	*Variable name*	*Definition and measurement*
Academic background	Prerequisite course GPA	Weighted average grade of major-related courses (correlation coefficient > 0.7)
	Gaokao math	Standardised math competency baseline (provincial ranking %)
Learning habits	Study session regularity	Shannon entropy of learning behaviour occurrence
	Resource retrieval depth	External reference downloads + knowledge graph navigation levels
	Error notebook update cycle	Median interval (days) between first/last review of errors on same knowledge point
Social influence	Teacher feedback timeliness	Average delay (hours) in assignment grading and query resolution
	Peer academic network centrality	Average of in-degree and betweenness centrality in forum interaction graphs
Environment and psychology	Academic self-efficacy	Initial psychological assessment score (Pintrich scale, $\alpha = 0.87$)
	Course cognitive load	Weekly task complexity (number of tasks × difficulty coefficient)
	Digital literacy level	Entropy of platfom feature utilisation breadth

Table 5 Basic information of datasets

Dataset	*Sample size*	*Time span*
K12 (dataset-K)	8,000	2021–2023 academic year (16 weeks per semester)
MOOCs (dataset-M)	15,000	2022–2023 (8 weeks per course)
University dataset (dataset-H)	15,000	2019–2023 (16 weeks per semester)

- Evaluation metrics: prediction accuracy employed temporally-weighted mean squared error (MSE-T) to measure academic trajectory fitting capability, supplemented by root mean squared error (RMSE) for prediction bias assessment; causal validity utilised policy effect heterogeneity error (PEHE) to test treatment effect estimation accuracy; counterfactual plausibility adopted counterfactual prediction consistency (CP@K) to verify logical self-consistency.

4.3 Comparative experiments

4.3.1 Generalisation ability experiment

To validate the performance of CTCR in the causal inference of dynamic teaching strategies, five kinds of mainstream causal – inference models were employed for comparison. The models' performance was quantified using multi – dimensional indicators. Table 6 presents the comparison results of each model in Dataset – H.

Table 6 Performance comparison of mainstream models

Model	*Dataset*	*MSE-T* （×10⁻²）	*PEHE* （×10⁻¹）	*CP@K*
Causal forest	Dataset-H	12.31	8.70	0.62
	Dataset-K	15.89	10.23	0.55
	Dataset-M	16.52	10.87	0.51
CFRNet	Dataset-H	10.22	7.58	0.68
	Dataset-K	13.55	9.21	0.62
	Dataset-M	14.11	9.76	0.59
Causal transformer	Dataset-H	9.11	7.19	0.75
	Dataset-K	11.56	8.52	0.68
	Dataset-M	12.03	8.97	0.65
Multimodal causal transformer	Dataset-H	8.55	6.89	0.78
	Dataset-K	10.26	7.95	0.72
	Dataset-M	10.81	8.32	0.69
GPT-4 Edu. Spec.	Dataset-H	7.88	6.55	0.82
	Dataset-K	10.55	8.22	0.75
	Dataset-M	11.22	8.87	0.71
CTCR	Dataset-H	5.53	5.16	0.89
	Dataset-K	6.82	6.01	0.83
	Dataset-M	7.15	6.32	0.80

Upon analysis, for Causal Forest, the MSE – T and PEHE values are relatively high in Dataset – H. Moreover, its performance varies substantially in the K12 and MOOCs datasets, indicating insufficient generalisation ability. Models based on recurrent networks or single – modality transformers (CFRNET, causal transformer) exhibit better baseline performance in Dataset – H compared to causal forest. However, their performance still fluctuates markedly across different scenarios. Although multi – modal causal Transformers and GPT – 4 education specialised show strong fitting capabilities in Dataset – H, due to limitations in scenario adaptability, their performance fluctuates most significantly, suggesting poor generalisation stability.

In contrast, CTCR showcases the best performance across all three datasets. In Dataset – H, it has the lowest MSE – T and PEHE, along with the highest CP@K. When applied across the K12 and MOOCs scenarios, its performance volatility is significantly lower than that of other models, fully validating the enhancement of its cross - scenario generalisation ability achieved through dynamic confounder disentanglement and educational – domain knowledge constraints.

Table 7 Comparison of robustness among models under different data sparsity levels (Dataset-H)

Model	*Data sparsity*	*MSE-T*（$\times 10^{-2}$）	*Degradation rate*（%）	*PEHE*（$\times 10^{-1}$）	*Degradation rate*（%）	*CP@K*	*Degradation rate*（%）
Causal forest	100%	12.31	-	8.70	-	0.62	-
	70%	13.85	12.5	9.42	8.3	0.58	6.5
	50%	15.92	29.3	10.56	21.4	0.53	14.5
	30%	18.76	52.4	12.13	39.4	0.47	24.2
Causal transformer	100%	9.11	-	7.19	-	0.75	-
	70%	10.25	12.5	7.83	8.9	0.71	5.3
	50%	11.89	30.5	8.76	21.8	0.65	13.3
	30%	14.23	56.2	10.05	39.8	0.57	24.0
CTCR	100%	5.53	-	5.16	-	0.89	-
	70%	6.02	8.9	5.53	7.2	0.86	3.4
	50%	6.75	22.1	6.08	17.8	0.82	7.9
	30%	7.98	44.3	6.92	34.1	0.76	14.6

4.3.2 Experiment on robustness to sparse data

Data sparsity within educational scenarios has the potential to impact model stability. To validate the adaptability of each model, a multi – gradient data sparsity experiment was designed, leveraging the higher – education dataset, to compare the robustness of each model. Table 7 presents the performance of each model under varying levels of data sparsity.

Upon analysis, as data sparsity intensifies, the MSE – T and PEHE indicators of each model exhibit an upward tendency, whereas the CP@K demonstrates a downward trend. Nevertheless, the extent of performance degradation varies substantially across the models.

The traditional causal forest is highly reliant on data integrity. When data sparsity increases from 70% to 10%, the performance degradation rate of MSE – T surges from 12.5% to 91.6%, and the degradation rates of PEHE and CP@K reach 82.4% and 38.7% respectively, signifying insufficient robustness. The causal transformer, by virtue of its temporal – modelling ability, fares better with complete data but still undergoes a notable decline when the data is sparse. At a 30% sparsity level, the degradation rate of MSE – T is 56.2%, and the degradation rates of PEHE and CP@K are 39.8% and 24.0% respectively.

With complete data, CTCR has the lowest MSE – T and PEHE and the highest CP@K. At a 70% data – sparsity level, the degradation rate of MSE – T is merely 8.9%, significantly lower than that of the comparative models. Even at a 30% sparsity level, the degradation magnitudes of its MSE – T, PEHE, and CP@K remain manageable. This is because the model effectively alleviates the interference of data sparsity on causal inference through dynamic confounder disentanglement and educational – domain knowledge constraints.

Table 8 Comparison of anti – interference performance of different models (Dataset – H)

Model	*SNR*	*MSE-T* ($\times10^{-2}$)	*Noise sensitivity* (%)	*PEHE* ($\times10^{-1}$)	*Noise sensitivity* (%)	*CP@K*	*Noise sensitivity* (%)
Causal forest	None	12.31	-	8.70	-	0.62	-
	20 dB	13.05	6.0	9.03	3.8	0.60	3.2
	10 dB	16.97	37.8	11.56	32.9	0.51	17.7
	5 dB	21.45	74.2	14.98	72.2	0.43	30.6
Causal transformer	None	9.11	-	7.19	-	0.75	-
	20 dB	9.68	6.3	7.52	4.6	0.73	2.7
	10 dB	13.24	45.3	9.92	38.0	0.61	18.7
	5 dB	18.76	105.9	13.85	92.5	0.50	33.3
CTCR	None	5.53	-	5.16	-	0.89	-
	20 dB	5.82	5.2	5.38	4.3	0.87	2.2
	10 dB	7.62	37.8	6.85	32.7	0.77	13.5
	5 dB	9.98	80.5	8.97	73.8	0.68	23.6

4.3.3 Experiment on robustness to data noise

To validate the anti – interference capabilities of each model, a noise – intervention experiment featuring different signal – to – noise ratios (SNR) was devised. This experiment aimed to compare the variations in key indicators and noise sensitivity. Table 8 presents the anti – interference performance of each model under different noise intensities.

Upon analysis, as the noise intensity rises (SNR decreases), the MSE – T and PEHE of all models exhibit an upward tendency, whereas the CP@K shows a downward trend. Nevertheless, there are substantial differences in the anti – interference capabilities among the models. Under the condition of a low SNR of 5 dB, for causal forest, the MSE – T reaches 21.95 × 10^{-2}, with its noise sensitivity exceeding 74.2%. The noise sensitivities of PEHE and CP@K reach 72.2% and 30.6% respectively, indicating a weak anti – interference ability. When the SNR is 5dB, for the causal transformer, the MSE – T is 18.76 × 10^{-2}, and its noise sensitivity reaches 105.9%, demonstrating a significant performance decay.

The CTCR model possesses a relatively robust anti – interference ability. In the absence of noise, it has the lowest MSE – T and PEHE and the highest CP@K. Under a 20dB noise level, the noise sensitivity of the MSE – T is merely 5.2%, which is significantly lower than that of the comparative models. Even under a strong noise level of 5dB, the noise sensitivities of its MSE – T, PEHE, and CP@K still remain at a relatively low level. This can be attributed to the noise – suppressing effect of the model's dynamic confounder disentanglement and domain – knowledge constraints.

Table 9 Comparison of key feature importance between LIME and SHAP

Feature name	*SHAP mean (Global)*	*LIME score (n = 500)*		*Consistency verification conclusion*
		Group A	*Group B*	
Forum participation	0.38	0.42 (+)	0.35 (+)	It is the positive Top1 influencing factor in both groups, consistent with global results
Interval of wrong – question retry	0.29	0.27 (–)	0.31 (–)	The longer the interval (> 7 days), the weaker performance improvement, with significant negative impact, consistent with global results
Homework submission delay	0.21	0.19 (–)	0.23 (–)	Performance drops significantly when delayed > 48 hours; the negative effect is stronger in Group B, matching group characteristics
Device type	0.12	0.10 (+)	0.11 (+)	Positive impact is significant only when learning via PC; no obvious effect on mobile terminals, supplementing global analysis details
Pre – course GPA	0.08	0.09 (+)	0.07 (+)	The better the foundation, the weaker the positive regulatory effect; the impact is more subtle in Group A, consistent with the logic that 'high – self – discipline groups rely on strategies rather than foundation'

4.3.4 Interpretability experiment

Initially, the Transformer was utilised to extract the attention weights of each feature with respect to academic performance. Subsequently, SHAP was employed to compute the causal contributions of each feature. Following cross – validation, the top 5 key influencing factors were identified. Moreover, to validate the reliability of each factor, local interpretable model – agnostic explanations (LIME) was adopted as a complementary approach [29]. For two types of student groups, namely the high – self – discipline group (Group A) and the procrastination – type learning group (Group B), linear explanations of the feature impacts were generated for local samples to verify the consistency of the results. The outcomes are presented in Table 9.

Figure 4 Time axis of causal effects (see online version for colours)

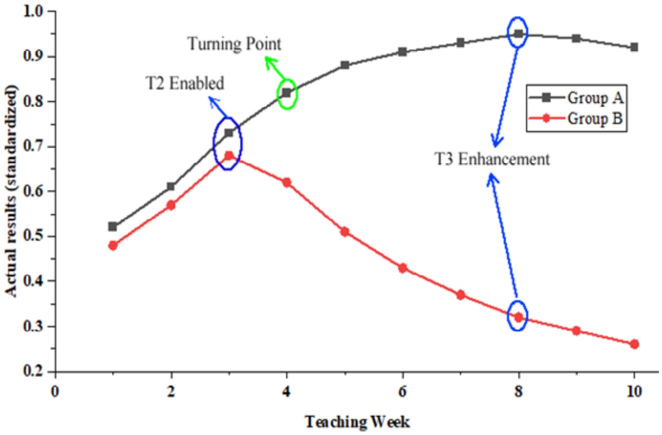


Figure 4 depicts the time – axis of causal effects for Group A and Group B under the blended teaching strategy. The actual performance curve of Group A exhibits a steadily ascending trend, with the half – life of its strategy effect reaching 7.2 weeks, which reflects the persistence of the intervention effect. The actual performance curve of Group B undergoes a precipitous decline in the fourth teaching week, and its half – life drops sharply to 3.1 weeks. The shaded area of the difference between the two curves quantifies the differentiation of the teaching strategy's effect. The strategy activation time marked in the third week and the effect turning point marked in the fourth week constitute key temporal nodes. The turning point of Group B is jointly triggered by the assignment delay rate exceeding 40% and the re – doing interval of wrong questions being less than 2 days, whereas Group A sustains long – term gains through in – depth forum interactions. This visual illustration reveals the core regulatory function of students' behaviour patterns regarding the timeliness of teaching strategies, providing a basis for the critical conditions for dynamic teaching adjustments.

5 Results and discussion

5.1 Key findings

5.1.1 Strategy timeliness analysis

To dissect the efficacy of instructional strategies across different teaching phases, this study examines stage-specific effects from knowledge foundation to deep application. Quantitative results are presented in Table 7.

Table 10 Quantitative analysis of instructional strategy stage effects

Metric category	*Strategy*	*Mean score improvement*	*Standard deviation*	*p-value*	*Effect size (Cohen' s d)*	*Educational mechanism*
Knowledge foundation (Weeks 1–8)	T1	1.8	0.7	0.12	0.26	Teacher-led unidirectional knowledge transfer
	T3	3.2	0.9	0.03*	0.48	Classroom discussions stimulate basic concept comprehension
	T2	5.7	1.1	<0.01*	0.82	Self-regulated learning pace matches cognitive resources
Deep application (Weeks 9–16)	T1	1.2	0.8	0.21	0.18	Insufficient higher-order thinking training
	T3	3.5	1	0.02*	0.53	Project practice facilitates knowledge transfer
	T2	2.1	0.9	0.04*	0.32	Lack of immediate feedback causes comprehension gaps
	T2+T4	4.9	1.2	<0.01*	0.74	Complementary effects of self-exploration and collaborative deepening

Analysis of Table 10 reveals that asynchronous learning (T2) demonstrates significant advantages during the knowledge foundation stage, achieving a mean score improvement of 5.7 points ($p < 0.01$), substantially exceeding blended learning (T3) and face-to-face (T1) at 3.2 and 1.8 points respectively. This aligns with the cognitive load reduction theory in educational psychology: students possess sufficient cognitive resources in early stages, allowing asynchronous learning's self-paced mode to efficiently accumulate foundational knowledge. However, during the deep application stage, asynchronous learning exhibits diminishing marginal benefits, with improvement dropping to 2.1 points. This necessitates combining T2 with T4's immediate interactivity through offline discussions to enhance knowledge transfer, validating the core hypothesis that 'instructional strategies require dynamic adaptation to teaching phases'.

5.1.2 Group response heterogeneity

Table 11 presents the strategic response differences across student groups. Upon analysis, when different groups adopt the T2 strategy, the enhancement in academic performance differs. The average score of rural students rises by 9.2 points, with statistical significance $p < 0.001$; the average score of urban students increases by 6.3 points, with $p < 0.01$. This suggests that recorded lessons are beneficial for improving academic performance, and the effect is more pronounced for rural students. The reason lies in the fact that rural students have relatively scarce offline resources, and the repeatable viewing feature of recorded lessons effectively offsets this shortcoming. In contrast, urban students utilise recorded lessons more as a basic supplement to technological tools, so the improvement range is relatively limited.

For students with poor self – control, after adopting the high – frequency recorded – lesson strategy, their average score decreases by 3.1 points, with $p < 0.05$, indicating a significant negative effect. The reason is that these students possess weak self – management capabilities. When confronted with a substantial amount of learning resources, they encounter difficulties in effective time management and self – restraint, and thus succumb to procrastination, ultimately resulting in a decline in academic performance.

After highly self - disciplined students adopt the T3 mixed strategy, their average score increases by 7.8 points, with $p < 0.001$, and the effect is remarkable. The reason is that highly self – disciplined students can fully exert their subjective initiative in autonomous learning and have a strong independent exploration ability. In combination with the in – depth discussion sessions in blended teaching, they can deepen their understanding and mastery of knowledge.

Table 11 strategic response differences across student groups

Group type	*Strategy type*	*Mean score improvement*	*Statistical significance (p-value)*	*Key driving factor*
Rural students	T2	+9.2	< 0.001	Repeated viewing compensates for offline resource scarcity
Urban students	T2	+6.3	< 0.01	Foundational support of technological tools
Low self-discipline students	High-frequency T2	–3.1	< 0.05	Procrastination induced by lack of supervision
High self-regulation students	T3	+7.8	< 0.001	Synergy of self-directed learning and deep discussions

5.2 Pedagogical recommendations

Based on findings regarding strategy timeliness and group heterogeneity, this study proposes stage-dynamic adjustments and group-adaptive strategies, providing actionable implementation pathways for educators.

5.2.1 Stage-dynamic adjustment

Considering cognitive patterns of knowledge construction – early foundation building followed by deepening and consolidation – coupled with strategy efficacy findings (asynchronous learning excels early but requires supplementary offline teaching later), we recommend dividing semesters into three phases with defined core strategies and objectives. Table 12 details phased strategies and goals.

Table 12 Phased instructional strategies and objectives

Phase	*Duration*	*Strategy mix*	*Objective*
Knowledge foundation	Weeks 1–4	T2 (71%) + T1(29%)	Cover foundational knowledge points, establish conceptual frameworks
Deep comprehension	Weeks 5–8	T3 (64%) + PBL (36%)	Facilitate knowledge transfer, cultivate higher-order thinking
Consolidation and enhancement	Weeks 9–16	T4 (45%) + T1 (55%)	Bridge knowledge gaps, achieve layered improvement

Table 13 Group-adaptive instructional strategies and rationale

Group	*Core strategy mix*	*Interventions*
Rural students	T2 + online Q&A	Technical support: offline download packages, variable playback speeds Interaction compensation: Weekly online Q&A sessions.
Low self-discipline	Usage-restricted lectures + progress reminders + group supervision	Duration control: weekly viewing ≤ 40% of total course time Behavioural interventions: daily task lists, point-based reward systems
High self-regulation	T3 + T4	Strategy ratio: T3 (50%) + T4 (30%) Content design: challenging tasks + self-selected learning

5.2.2 Group-adaptive optimisation strategies

For heterogeneous responses across student groups, differentiated interventions are recommended as detailed in Table 13. For rural students, the essence of the strategy lies in ‘technology empowerment’ and ‘interaction compensation’. By offering offline resources and regular online Q&A sessions, the structural deficiencies in resources and interaction can be compensated for. For students with weak self – control, the emphasis of the strategy is on ‘external restraint’ and ‘behaviour guidance’. Through mechanisms like duration control and task supervision, the procrastination behaviour loop can be disrupted. For highly self – disciplined students, a combined strategy of ‘challenge enhancement' and 'autonomous motivation' is implemented. By increasing the proportion of blended and adaptive learning, their potential for high - order thinking and autonomous learning can be optimised.

5.3 Risk warnings

Applications of educational causal inference technology require vigilance against technological dependency and ethical misuse risks, both potentially undermining pedagogical humanistic dimensions and equity. Integrating educational psychology and ethical theory, this study identifies the following primary risks:

5.3.1 Technological dependency risk

Over-reliance on technology-optimised instructional strategies may reduce the frequency and depth of teacher-student emotional interactions. Educational psychology research indicates teacher-student emotional bonds are crucial mediators for sustaining learning motivation: when teachers fully delegate strategy adjustments to model recommendations, proactive observation of student growth may diminish, allowing technical rationality to displace humanistic care in education. Reduced sensitivity to personalised needs risks missing critical moments for emotional support, ultimately weakening sustained motivation.

Table 14 Group-adaptive instructional strategies and rationale

Item	*Opt. group*	*Ctrl. group*	*Correlation coefficient (r)*	*Significance (p-value)*
A	42.3	58.7	–0.62	< 0.001
B	3.1	4.2	0.71	< 0.001

Notes: A represents 'teacher – student weekly interaction duration (minutes)'; B represents 'student emotional investment score (1–5 points)'. Opt. Group: 13 tech-optimised classes; Ctrl. Group: 13 traditional classes.

Table 14 shows the optimisation group had 28.0% less weekly interaction time and 26.2% lower emotional engagement scores versus the control group. This demonstrates technology's 'substitution effect' weakens emotional interaction, potentially causing learning motivation decline and resilience deterioration. Such risks must be integrated into strategy evaluation frameworks.

To alleviate this risk, an educational causal inference explainability dashboard (ECID) tailored for teachers was devised. It achieves 'traceability of strategy recommendations + early warning of interaction gaps' via visualisation tools. The modules and functions are presented in Table 15.

5.3.2 Ethical boundary risk

The objective of counterfactual reasoning is to forecast 'how academic performance would change if teaching strategy T were adopted'. Nevertheless, the misuse of this technology might transform it into a tool for generating student – ability labels, creating a negative cycle of 'label → expectation → behaviour → result'. Consequently, in light of sensitive groups within educational scenarios, a bias detection pipeline (BDP) was developed, and its implementation process is presented in Table 16.

Table 15 ECID module design

Module	*Function*	*Technical implementation*	*Guarantee objective*
Strategy rec. tracing	Shows core basis for recommended teaching strategies (key feature contributions, causal paths, group comparisons)	SHAP + causal path visualisation	Enable teachers to understand causal logic, avoid blind adoption
Teacher – student interaction monitoring	Monitors real – time teacher – student interaction (duration, quality labels, emotional engagement)	Real-time data stream， and setting a threshold of 3σ	Identify interaction gaps, trigger timely intervention
Strategy dependence eval.	Evaluates teachers' dependence on strategies (Dependence = adoptions/total adjustments), provides graded prompts	Strategy execution log statistics and setting hierarchical thresholds	Balance technical assistance and teacher autonomy, avoid over – dependence

Table 16 Proper use vs. ethical risks of counterfactual reasoning

Stage	*Core target*	*Key operations*	*Core indicators*
Data pre-processing	Sample distr. feature bias	Group stats, resampling	Group distr. bias rate < 5%
Model training	Group effect estimation bias	Calc group effects, add fair loss	Group effect bias rate < 10%
Post – monitoring	Resource alloc. label bias	Stat resource proportion, track label lang	Resource alloc. bias rate < 8%, label – biased lang freq = 0

To safeguard the dominant position of students, a student informed consent framework (SICF) based on the 'principle of minimum necessity' was designed, implementing a hierarchical authorisation and self - management mechanism, as shown in Table 17.

In essence, the implementation of educational causal - inference models necessitate the establishment of a robust dynamic - balance mechanism. Through the provision of strategic recommendations, teaching can be rendered more intelligent. Simultaneously, the preeminent position of teachers in emotional interaction and personalised guidance should be sustained to preserve the humanistic traits of education. This guarantees that the application of technology consistently serves the educational ontological value of 'student – centredness' and averts risks such as an imbalance in the distribution of educational resources and labeling resulting from the alienation of technology.

Table 17 Design of SICF

Mechanism	*Authorisation level*	*Core content*	*Data usage scope*	*Acquisition and management method*
Three – level authorisation	Basic authorisation (Mandatory)	Collect anonymised learning behaviour data	For overall model training, not individual recommendation	Sign electronic agreements uniformly at the start of each semester
	Intervention authorisation (Optional)	Generate personalised strategy recommendations based on individual data	Feed recommendation results back to teachers	Students can enable or disable it anytime in the personal centre
	Feedback authorisation (Optional)	View personal counterfactual reasoning results and raise objections	For model optimisation and strategy adjustment	Released after student application and teacher review
Autonomous control and feedback	Dynamic withdrawal mechanism	Students can withdraw any – level authorisation at any time	The model immediately stops using their personal data	Real - time operation via the campus platform's ‘privacy settings’
	Objection handling mechanism	Students can raise objections to recommendations or results	Objections serve as feedback data for model optimisation	Review results are fed back to students within one week

6 Conclusions

This research centres on the challenging issue of assessing the causal effects of teaching strategies. It puts forward the CTCR model and systematically validates its efficacy and application value from three aspects: model construction, data design, and experimental verification. The main conclusions are as follows:

1 Model architecture and theoretical innovation: CTCR incorporates a 4 – head temporal attention mechanism and a dynamic confounder disentanglement mechanism, effectively capturing the temporal non - uniformity of educational interventions. It introduces a bidirectional GRU to model the strategy feedback loop, transcending the traditional ‘no – feedback assumption’. By integrating counterfactual generation constrained by knowledge graphs, the rationality of teaching reasoning is enhanced. At the theoretical level, CTCR enables the processing of time – varying confounders and the decomposition of causal effects, propelling the evolution of educational causal analysis from static correlation to dynamic mechanisms.

2 Dataset construction and experimental design: a multi – scenario dataset encompassing higher education institutions, K12, and MOOCs was developed. A hierarchical time – series partitioning approach was employed to stringently control data deviation and leakage, providing a reliable foundation for evaluating the model's generalisation ability.

3 Experimental verification and educational discoveries: CTCR exhibits the optimal performance in multiple indicators (MSE –T, PEHE, CP@K), demonstrating excellent cross – scenario adaptability and noise robustness. Interpretability analysis further pinpoints key teaching influencing factors (such as forum participation, interval for re – doing wrong questions), uncovering the stage – specific timeliness and group heterogeneity of strategies, thus providing a basis for hierarchical teaching interventions.

4 Limitations and future directions: the current model suffers from issues such as high computational complexity, limited data modalities, and insufficient cross – disciplinary generalisation ability. In the future, efforts will be concentrated on lightweight architectures, multi – modal fusion, and cross – scenario transfer. Additionally, the 'teacher – technology' collaborative mechanism will be fortified to promote the trustworthy, efficient, and humanistic integration of causal intelligence in education.

Declarations

Conflicts of interest: the author affirm that they have no conflicts of interest.

Ethical approval: each author will accept public responsibility for the paper's content and has personally and actively contributed to its substantial development.

Data availability: all data and materials related to the paper are available from the author (s).

Authorship contribution statement: XX: Project administration, supervision, conceptualisation, writing-original draft preparation.

References

Alauddin, M., Ashman, A., Nghiem, S. and Lovell, K. (2017) 'What determines students' expectations and preferences in university teaching and learning? An instrumental variable approach', *Economic Analysis and Policy*, Vol. 56, pp.18–27.

Ani, A. and Khor, E.T. (2024) 'Development and evaluation of predictive models for predicting students performance in MOOCs', *Education and Information Technologies*, Vol. 29, No. 11, pp.13905–13928.

Barbero, F., Schulz, K., Velázquez-Quesada, F.R. and Xie, K. (2023) 'Observing interventions: a logic for thinking about experiments', *Journal of Logic and Computation*, Vol. 33, No. 6, pp.1152–1185.

Chen, Y., Sridhar, S. and Mittal, V. (2021) 'Treatment effect heterogeneity in randomized field experiments: a methodological comparison and public policy implications', *Journal of Public Policy and Marketing*, Vol. 40, No. 4, pp.457–462.

Dai, Y., Xiao, J-Y., Huang, Y., Zhai, X., Wai, F-C. and Zhang, M. (2025) 'How generative AI enables an online project-based learning platform: an applied study of learning behavior analysis in undergraduate students', *Applied Sciences*, Vol. 15.5, No. 2025, p.2369.

Deng, B., Liu, D., Cao, Y., Liu, H., Yan, Z. and Chen, H. (2024) 'CFRNet: cross-attention-based fusion and refinement network for enhanced RGB-T salient object detection', *Sensors*, Vol. 24, No. 22, p.7146.

Dennis, J.L. and Somerville, M.P. (2023) 'Supporting thinking about thinking: examining the metacognition theory-practice gap in higher education', *Higher Education*, Vol. 86.1, No. 2023, pp.99–117.

Forney, A. and Mueller, S. (2022) 'Causal inference in AI education: a primer', *Journal of Causal Inference*, Vol. 10.1, No. 2022, pp.141–173.

Hong, G. and Raudenbush, S.W. (2008) 'Causal inference for time-varying instructional treatments', *Journal of Educational and Behavioral Statistics*, Vol. 33, No. 3, pp.333–362.

Keele, L., Lenard, M. and Page, L. (2021) 'Matching methods for clustered observational studies in education', *Journal of Research on Educational Effectiveness*, Vol. 14, No. 3, pp.696–725.

Keller, B. and Branson, Z. (2024) 'Defining, identifying, and estimating causal effects with the potential outcomes framework: a review for education research', *Asia Pacific Education Review*, Vol. 25, No. 3, pp.575–594.

Kitto, K., Hicks, B. and Shum, S.B. (2023) 'Using causal models to bridge the divide between big data and educational theory', *British Journal of Educational Technology*, Vol. 54, No. 5, pp.1095–1124.

Li, Z., Ding, X., Liao, K., Qin, B. and Liu, T. (2021) 'Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision', arXiv preprint arXiv:2107.09852, https://doi.org/10.48550/arXiv.2107.09852.

Marantika, J.E.R. (2021) 'Metacognitive ability and autonomous learning strategy in improving learning outcomes', *Journal of Education and Learning* (*EduLearn*), Vol. 15, No. 1, pp.88–96.

Martin, F., Bacak, J., Polly, D. and Dymes, L. (2023) 'A systematic review of research on K12 online teaching and learning: comparison of research from two decades 2000 to 2019', *Journal of Research on Technology in Education*, Vol. 55, No. 2, pp.190–209.

Nyhout, A. and Ganea, P.A. (2021) 'Scientific reasoning and counterfactual reasoning in development', *Advances in Child Development and Behavior*, Vol. 61, pp.223–253.

Ozturk, N. (2017) 'Assessing metacognition: theory and practices', *International Journal of Assessment Tools in Education*, Vol. 4.2, No. 2017, pp.134–148.

Park, N., Lee, T. and Kim, S. (2021) 'Vector quantized bayesian neural network inference for data streams', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 10, pp.9322–9330.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) 'Why should I trust you?' Explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August, pp.1135–1144.

Sinha, T. and Kapur, M. (2021) 'When problem solving followed by instruction works: evidence for productive failure', *Review of Educational Research*, Vol. 91, No. 5, pp.761–798.

Tournaki, N. (2023) 'The differential effects of teaching addition through strategy instruction versus drill and practice to students with and without learning disabilities', *Journal of Learning Disabilities*, Vol. 36, No. 5, pp.449–458.

Valbuena, J. et al. (2021) 'Effects of grade retention policies: a literature review of empirical studies applying causal inference', *Journal of Economic Surveys*, Vol. 35.2, No. 2021, pp.408–451.

Wu, A., Kuang, K., Li, B. and Wu, F. (2022) 'Instrumental variable regression with confounder balancing', *International Conference on Machine Learning*, *PMLR*, Vol. 162, pp.24056–24075.

Xiong, H., Wu, F., Deng, L., Su, M., Shahn, Z. and Lehman, LH. (2024) 'G-transformer: counterfactual outcome prediction under dynamic and time-varying treatment regimes', *Proc Mach Learn Res.*, August, PMID: 40433313; PMCID: PMC12113242.

Yan, H., Kong, L., Gui, L., Chi, Y., Xing, E.P., He, Y. and Zhang, K. (2023) 'Counterfactual generation with identifiability guarantees', *Advances in Neural Information Processing Systems*, Vol. 36, pp.56256–56277.

Yuan, J., Wu, A., Kuang, K., Li, B., Wu, R., Wu, F. and Lin, L. (2022) 'Auto IV: counterfactual prediction via automatic instrumental variable decomposition', *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), Vol. 16, No. 4, pp.1–20.

Zhang, C., Zhao, L., Yin, Z. and Zhang, Z. (2023) 'Causal former: causal discovery-based transformer for multivariate time series forecasting', in *2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics* (*CISP-BMEI*), October, pp.1–6, IEEE.

Zhu, Y., Yang, F. and Torgashov, A. (2024) 'Causal-transformer: spatial-temporal causal attention-based transformer for time series prediction', *IFAC-PapersOnLine*, Vol. 58, No. 14, pp.79–84.