Invariant Representation Learning for Memory Behavior Modeling via Adaptive Environment Separation

Xiaoxuan Shen[1], Zhihai Hu[1], Fuqing Li[2], Shengyingjie Liu[1], Jianwen Sun[1]*

[1]National Engineering Research Center for Educational Big Data, Central China Normal University
[2]Institute of Collaborative Innovation, University of Macau
{shenxiaoxuan,sunjw}@ccnu.edu.cn,{huzhihai,lsyj}@ccnu.mails.edu.cn, mc46655@um.edu.mo

Abstract

Memory behavior modeling seeks to predict individual recall performance and understand its underlying cognitive mechanisms. However, the dynamic and heterogeneous nature of memory data poses significant challenges to the generalization ability of models under unseen conditions. To address this challenge, we propose an invariant representation learning framework I-Mem that integrates self-supervised contrastive learning with decorrelation constraints, enabling the adaptive identification and suppression of environment-related factors in sequential behavioral data, thereby mitigating the influence of spurious features and enhancing the modeling of stable cognitive structures. Importantly, the method does not rely on explicit environment partitioning or predefined environment labels, while our theoretical analysis demonstrates that it can effectively resist environmental perturbations and facilitate the extraction of invariant structural representations, thereby ensuring adaptability and generalization. Empirical evaluations on both synthetic and real-world datasets further confirm its superiority over mainstream methods in terms of generalization performance and stable feature identification. Feature attribution analysis reveals that I-Mem extracts invariant features aligned with classical cognitive effects, and reflects short-term behavioral patterns that may indicate latent cognitive mechanisms beyond existing theories, highlighting both interpretability and discovery potential.

Code — https://github.com/hellowads/I-Mem

Introduction

Memory behavior modeling is a fundamental task in cognitive science and educational technology, aiming to understand and predict individual memory performance over time(Shen et al. 2026). It supports a variety of applications, such as learning outcome assessment (Shen et al. 2025), and personalized instructional strategy design(Walkington 2013). Traditional research has proposed a series of explanatory models grounded in classical cognitive theory, such as the Ebbinghaus forgetting curve(Ebbinghaus 2013) and ACT-R(Anderson et al. 2004). These models offer strong theoretical interpretability under controlled experimental conditions and have contributed significantly to our understanding of memory mechanisms in psychology. However, due to their reliance on small-scale, controlled datasets, these models struggle to capture individual variability and temporal dynamics, leading to poor generalization in real-world conditions(Walsh et al. 2018). While memory behavior modeling aspires to uncover generalizable regularities in recall behavior, yet such regularities remain elusive—potentially due to the limitations of current methods in extracting them under dynamic conditions.

In recent years, deep learning models built on large-scale behavioral data from platforms like Anki and Duolingo have emerged(Pearlin and Gandhi 2024; Ye, Su, and Cao 2022). Nevertheless, these models still rely on empirical risk minimization under the assumption of in-distribution data (Vapnik 1999), lacking structured mechanisms for handling out-of-distribution scenarios. In other words, existing methods exhibit a blind spot in their modeling perspective—they overlook the critical challenge of generalization. This difficulty may stem not from insufficient model complexity, but from a misaligned modeling perspective: current approaches tend to optimize performance within fixed experimental settings, rather than extract stable cognitive structures from heterogeneous, dynamic behavioral data(Carvalho and Lampinen 2025).

Invariant learning (IL) offers a theoretical foundation to tackle this problem by focusing on predictive structures that remain stable across environments(Ye et al. 2024). While existing methods such as IRM(Arjovsky et al. 2019), GroupDRO(Sagawa et al. 2019), and FOIL(Liu et al. 2024a) have made progress in other domains, their applicability to memory behavior modeling remains limited. This is largely because memory behavior exhibits continuous, implicit changes over time, such as those caused by repeated reviews or external cues. These subtle and dynamic shifts make it difficult to segment memory data into discrete environments, thus violating the core assumptions underpinning existing invariant learning approaches(Mozer and Lindsey 2016).

To address these challenges, we propose a novel invariant learning framework I-Mem tailored for sequential memory behavior modeling. Unlike traditional approaches that rely on static environment partitioning and fixed causal assumptions, I-Mem explicitly models and removes spurious cor-

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

relations. Specifically, it introduces a self-supervised contrastive learning strategy combined with a decorrelation objective, which adaptively identifies and suppresses unstable, spurious features induced by environmental variations in the representation space. Without requiring explicit environment labels or predefined partitions, I-Mem directly operates on dynamic and heterogeneous memory behavior data, offering a new perspective for building generalizable models in real-world learning scenarios. The main contributions of this work are:

- An invariant representation learning framework is proposed, which requires no environment labels and is directly applicable to time-series memory behavior data, making it well-suited for modeling under continuous and implicit distribution shifts.
- A flexible and generalizable feature disentanglement mechanism is designed to adaptively identify and suppress spurious features in the representation space, enhancing the modeling of stable cognitive patterns without relying on explicit environment partitioning or prior knowledge.
- Comprehensive empirical evaluations are performed on both real-world and synthetic memory behavior datasets, and combined with feature importance analysis, I-Mem is validated to effectively identify stable features, reduce spurious dependencies, and improve out-of-distribution generalization performance.
- The invariant features exhibit trends broadly consistent with classical memory theories (e.g., memory decay), while also uncovering novel short-term behavioral patterns (e.g., transient recall variability), offering new directions for cognitive modeling.

Background

Memory Behavior Modeling: Early studies on memory behavior modeling primarily focused on using mathematical models to explain and predict human memory behavior(Anderson and Schooler 1991). These models were often built upon controlled experimental paradigms, where memory performance metrics (e.g., recall rate R) were linked to behavioral features through functional relationships. The earliest model dates back to Ebbinghaus's forgetting curve(Ebbinghaus 2013) proposed in 1885. Subsequent works introduced a variety of memory equations, such as Wickelgren's generalized power law model(Wickelgren 1974), Anderson's ACT-R model(Anderson et al. 2004), and the Half-Life Regression (HLR) model(Settles and Meeder 2016). While these models are simple and interpretable, they often suffer from challenges in parameter quantification. For instance, the forgetting rate in Wickelgren's formula is not directly observable and lacks a clear mapping to observable behavioral features(Wixted 2004).

With the availability of large-scale behavioral data, neural architectures such as RNNs(Piech et al. 2015) and Transformers(Liu et al. 2023) have been adopted for memory behavior modeling(Ma et al. 2023), significantly improving predictive accuracy. These models are typically trained under empirical risk minimization and rely on the i.i.d. assumption, yielding strong in-distribution performance. However, real-world learning behavior is shaped by individual differences, strategy shifts, and task variability, leading to dynamic distribution shifts and environmental heterogeneity(Liu et al. 2024b). As a result, model performance deteriorates sharply in out-of-distribution scenarios(Bayat et al. 2024). From a theoretical perspective, memory behavior modeling seeks to uncover generalizable cognitive principles beyond behavioral pattern fitting(Goyal and Bengio 2022). However, poor out-of-distribution robustness often prevents current methods from capturing such stable structures, limiting both interpretability (Sun et al. 2025; Liu et al. 2024c) and generalization(Bayat et al. 2024).

Time-series invariant learning: Time-series invariant learning aims to extract stable and generalization representations across varying environments(Wu et al. 2025). Prior approaches include environment-based invariance constraints(Arjovsky et al. 2019), adversarial alignment(Ganin and Lempitsky 2015), causal modeling(Schölkopf et al. 2021), and self-supervised contrastive learning(Chen et al. 2020). While effective in domains such as vision, these methods face key limitations in memory behavior modeling. First, the feature-label relationship evolves over time, making spurious correlations more likely to dominate(Liu et al. 2024b; Lu et al. 2021). Second, distribution shifts and environmental changes are typically continuous and implicit, rendering static environment partitioning ineffective(Liu et al. 2021). To address these challenges, we propose an invariant representation learning framework that does not rely on environment labels. This design is motivated by a key observation: the widely adopted InfoMax principle introduces an inductive bias in sequential modeling by amplifying all predictive signals, including spurious and non-causal features. Such amplification impairs the model's ability to capture stable structures under out-of-distribution conditions. Building on this insight, I-Mem leverages augmentation-induced sensitivity to spurious features to infer latent invariant structures, and introduces a feature-level disentanglement mechanism to suppress spurious features. This enables robust representation learning in dynamic and heterogeneous memory behavior sequences.

Preliminaries

Notation. Throughout this paper, we use C and S to denote the invariant and spurious factors, respectively. These are interchangeable with the invariant representation h_c and the spurious representation h_s. The estimated factors are denoted as $\hat{C}$ and $\hat{S}$, and their corresponding representations as $\hat{h}_c$ and $\hat{h}_s$. $g(\cdot)$ to denote a function. The detailed notation can be found in supplementary material.

Problem Definition. In memory behavior modeling, the input memory time-series is denoted as $X \in \mathbb{R}^{l \times F}$, where l is the length of the observation window and F is the dimensionality of memory features at each time step. The prediction target is $Y \in \mathbb{R}^{l \times d_{\text{out}}}$, where d_{out} is the dimensionality of the prediction target for each step. At time step t, a training sample is represented as (X, Y), where $X = [x_{t-l+1}, x_{t-l+2}, \ldots, x_t]$ and $Y = [y_t, y_{t+1}, \ldots, y_{t+l}]$.

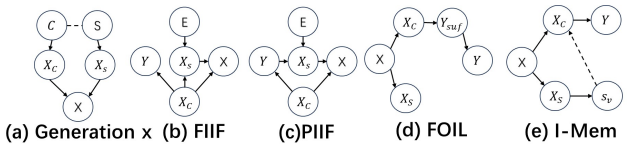

Figure 1: Overview of causal assumptions and methodological differences. (a) x is composed of features generated from invariant factors C and spurious factors S. (b) fully informative invariant feature (FIIF): The label Y is determined solely by invariant factors C; (c) partially informative invariant feature (PIIF): Spurious factor S also influences Y, introducing spurious correlation; (d) FOIL serves as a representative example, which facilitates invariant learning by inferring latent environment labels; (e) I-Mem: We explicitly isolate spurious features S using cognitive disentanglement to enhance the learning of invariant representations C, improving generalization in OOD scenarios.

We focus on the Out-of-Distribution (OOD) generalization problem in memory sequence prediction. Given multiple heterogeneous memory datasets $\mathcal{X} = \{X^e\}_{e \in \varepsilon_e \subseteq \varepsilon_{\text{all}}}$, the goal is to train a time-series prediction model g_θ, parameterized as $g_\theta = \rho \circ h$, where the encoder $h : X^e \to \mathbb{R}^d$ learns a memory representation, and the classifier $\rho : \mathbb{R}^d \to Y$ predicts the output as $\hat{Y}_X = \rho(h(X^e))$. The objective is to learn an optimal model $g(\cdot)$ that performs well across all seen and unseen environments:

$$g(\cdot) = \arg\min_{g} \sup_{e \in e_{\text{all}}} \mathcal{R}(g \mid e) \tag{1}$$

where $\mathcal{R}(g \mid e) = \mathbb{E}^e_{(X^e,Y)}[\ell(f(X^e), Y)]$ denotes the risk of model g in environment $e \in e_{\text{all}}$, and $\ell(\cdot, \cdot)$ is the loss function. To optimize Equation 1, Invariant Learning(IL) aims to identify and exploit invariant features—those that maintain a stable relationship with the target variable across environments. For instance, in memory behavior modeling, the forgetting rate of a learner is often an invariant feature, whereas the identity of flashcards may represent environment-specific (variant) factors.

The Framework of the I-Mem Method

In memory behavior modeling, InfoMax-based representations often overfit spurious features due to inductive biases in sequential data. Motivated by this observation, we propose I-Mem, an invariant representation learning framework tailored for dynamic, heterogeneous time-series data.

Spurious Feature Bias in InfoMax Representations

In this section, we analyze how self-supervised learning (SSL), particularly InfoMax-based methods, can effectively identify spurious features in time-series data without relying on external labels.

$$\max_{\theta} \frac{1}{|\mathcal{X}||X||x|} \sum_{X^e \in \mathcal{X}} \sum_{x_t^e \in X^e} \sum_{f_{t,i}^e \in x_t^e} I(\hat{h}_f; \hat{h}_x) \tag{2}$$

where $\hat{h}_f$ represents the representation of the i-th local feature at time step x_t, and $\hat{h}_x$ denotes the global representation at time step x_t, e represents different memory sequences θ denotes the model parameters. T represents the full temporal sequence. Equation 2 aims to maximize the mutual information between $\hat{h}_f$ and $\hat{h}_x$, intuitively encouraging the model to capture information patterns across all feature in the input sequence. However, we observe that the global representation $\hat{h}_x$ tends to be biased toward spurious features. Formally, The spurious representation $\hat{h}_s$ can be characterized as:

$$\hat{h}_s = \max_{\theta} \frac{1}{|\mathcal{X}||X||x|} \sum_{X^e \in \mathcal{X}} \sum_{x_t^e \in X^e} \sum_{f_{t,i}^e \in x_t^e} I(\hat{h}_f; \hat{h}_x) \tag{3}$$

Theorem 1. *Let C denote the invariant factor and S the spurious factor in a time-series data generation process. Assume the Shannon entropies satisfy $H(S) \geq H(C)$, and each local representation $\hat{h}_f$ contains bounded mutual information about both factors, such that: $\delta_r \geq I(\hat{h}_f; C) - I(\hat{h}_f; S) \geq \delta_l, \quad \forall f \in x_c$; $\delta'_r \geq I(\hat{h}_f; S) - I(\hat{h}_f; C) \geq \delta'_l, \quad \forall f \in x_s$. If the ratio $|x_s| \cdot \delta'_l > |x_c| \cdot \delta_r$, then the InfoMax objective will predominantly encode spurious features S.*

Theoretical proof and empirical validation are provided in supplementary material. Although the condition $|x_s| \cdot \delta'_l > |x_c| \cdot \delta_r$ is not directly verified, it often holds in practice, as spurious features are typically more numerous and variable than invariant ones (Lu et al. 2021). For example, in memory behavior modeling, models may overfit to environment-driven factors such as task types, while neglecting stable memory mechanisms. This inductive bias explains InfoMax's tendency to capture spurious correlations and motivates our emphasis on feature disentanglement.

The I-Mem Framework

Having discussed how InfoMax-based representations $\hat{h}_s$ tend to encode spurious features , we propose I-Mem, a three-stage framework—Encoding, Mapping, and Decorrelation—that isolates spurious components from InfoMax representations and learns invariant features with OOD generalization. The overall workflow is illustrated in Figure 2, and the learning objective is defined as follows:

$$\max I(\hat{h}_c; Y), \quad \text{s.t. } \hat{h}_c \perp E,\ \hat{h}_c = h(X) \tag{4}$$

Stage 1: Encoding. In the first stage, we train an encoder using the InfoMax objective in Equation 2, which typically produces representations entangled with spurious components $\hat{h}_s$. However, a single encoder may only capture a subset of spurious features, limiting downstream decorrelation. To address this, we construct a potential representation set $\mathcal{H}$ from multiple augmentation views to comprehensively cover the distribution of spurious features.

Stage 2: Mapping. Given the representation set $\mathcal{H}$. To enforce $\hat{h}_c \perp e$, we instead aim to maximize the conditional entropy:

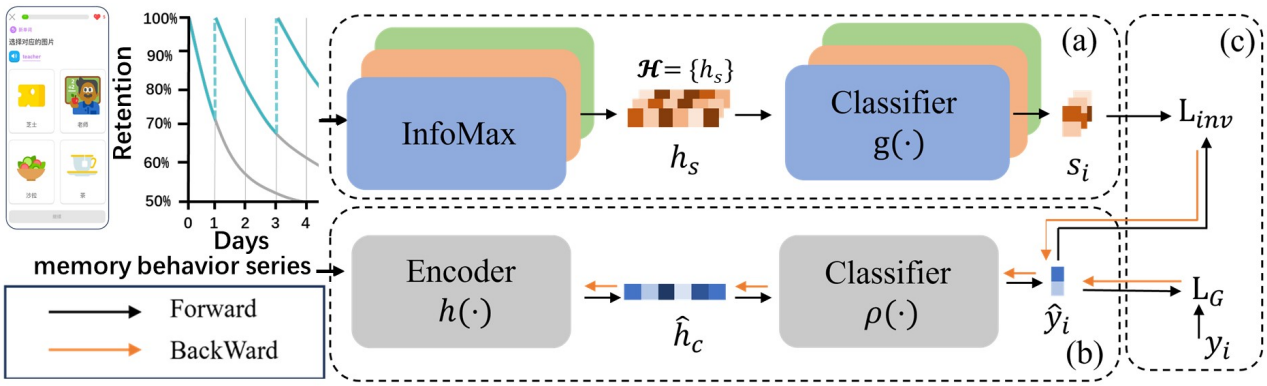

Figure 2: Model Architecture.(a) Encoding and Mapping: The model first employs an InfoMax-based encoder to extract environment-sensitive representations. These are then processed through a mapping module to generate a candidate set of spurious environment-specific representations, denoted as $\mathcal{H}$.(b+c) Decorrelation: A new model is retrained using a decorrelation objective that explicitly removes spurious information from the learned representation by leveraging $\mathcal{H}$. This enhances the invariance and robustness of the final representation under distributional shifts.

$$\max H(\hat{h}_s \mid \hat{h}_c) \tag{5}$$

However, since $\hat{h}_s$ is high-dimensional, directly optimizing Eq. 5 is intractable. As an alternative, we use the potential representation set $\mathcal{H}$ as input to train multiple independent classifiers $\{g_i\}$ with the label Y as the supervision target. Given that InfoMax representations predominantly encode spurious components, these classifiers are expected to capture correlations between spurious patterns and the labels. The output of each classifier s_i is then regarded as a compact approximation of the spurious representation $\hat{h}_s$.

Stage 3: Decorrelation. After obtaining the pseudo-label set $\mathcal{S} = \{s_i\}$, we design a decorrelation objective to explicitly remove spurious information from the final representation $\hat{h}_c$. Specifically, we construct a classifier ρ to predict the $\hat{y}_i = \rho(\hat{h}_c)$, and maximize the mutual information between the $\hat{h}_c$ and the pseudo-labels s_i through a decorrelation loss. To improve training stability under label imbalance, we apply a normalization to the output logits and formulate the loss as a conditional entropy maximization surrogate:

$$\mathcal{L}_{\text{Inv}} = \max_{\rho} \left(-\frac{1}{N} \sum_i s_i \log \rho(\hat{h}_c) \right) \tag{6}$$

To theoretically support the effectiveness of this decorrelation mechanism, we propose the following corollary:

Corollary 1. *If the pseudo-label s_i is generated from the spurious factor S, and satisfies $I(s_i; S) \geq \gamma > 0$, then maximizing the decorrelation loss $\mathcal{L}_{Inv}$, which minimizes $I(\hat{h}_c; s_i)$, leads to $I(\hat{h}_c; S) \to 0$, thereby significantly improving the generalization ability of the learned representation.*

The proof of Corollary 1 is provided in supplementary material. Finally, we combine the decorrelation loss with the supervised task loss to form the overall training objective:

$$\mathcal{L} = \mathcal{L}_{\text{G}} - \lambda \mathcal{L}_{\text{Inv}} \tag{7}$$

where λ controls the strength of the decorrelation loss and $\mathcal{L}_{\text{G}}$ is the standard supervised loss.In summary, our algorithm consists of three key stages: (1) an InfoMax-based encoder is trained to extract environment-sensitive representations; (2) The mapping module uses multi-view prediction to expose spurious feature correlations; (3) a decorrelation loss is applied to suppress these components and isolate invariant representations.

Experiments

In this section, we conduct extensive experiments to address the following research questions:

- **RQ1:** Does I-Mem achieve better or comparable generalization performance compared to state-of-the-art baselines?
- **RQ2:** To what extent do the key components and hyperparameters of I-Mem influence the final performance?
- **RQ3:** How can we verify that I-Mem indeed learns invariant features with OOD generalization capability through spurious feature decorrelation?
- **RQ4:** In real-world memory scenarios, can we systematically identify useful invariant features that offer insights for cognitive modeling?

Experimental Setup

Synthetic Data. To validate the model's ability to distinguish invariant from spurious factors under controlled conditions, we construct synthetic datasets that simulate memory behavior based on the Wickelgren forgetting function: $R = \lambda(1 + \beta t)^{-\psi}$. Each sequence X comprises invariant features X_C and spurious features X_S, governed by latent variables $C = (\beta, \lambda, \psi)$ and S, respectively. The label Y depends on C, while S varies across environments and may introduce spurious correlations. We consider two settings—FIIF (Wu et al. 2022) and PIIF (Chen et al. 2023)—to model different invariance conditions (see Figure 1). Details are provided in supplementary material.

Model	FIIF setting						PIIF setting					
	α =0.3		α =0.6		α =0.9		(α,β)=(0.7,0.8)		(α,β)=(0.9,0.8)		(α,β)=(0.9,0.7)	
	MAE ↓	AUC ↑	MAE ↓	AUC ↑	MAE ↓	AUC ↑	MAE ↓	AUC ↑	MAE ↓	AUC ↑	MAE ↓	AUC ↑
ERM	0.4945	0.6285	0.3805	0.6843	0.0950	0.9355	0.5225	0.6126	0.5065	0.7236	0.4681	0.6954
GroupDRO	0.4991	0.6135	0.3746	0.6832	0.1108	0.9101	0.5315	0.6094	0.4361	0.7415	0.3706	0.7033
IRM	0.5245	0.5899	0.4262	0.6613	0.1277	0.8463	0.5128	0.6005	0.5163	0.7091	0.5034	0.6362
IB-IRM	0.5348	0.5416	0.4401	0.6277	0.2312	0.8235	0.5281	0.6038	0.5652	0.7153	0.5568	0.6701
EIIL	0.5348	0.5902	0.5382	0.6684	0.5063	0.8792	0.5281	0.5851	0.5035	0.7191	0.5191	0.6694
TKNet	0.4934	0.5882	0.4622	0.6928	0.4634	0.9132	0.5146	0.5514	0.4345	0.7146	0.4423	0.6329
FOIL	0.4966	0.6214	0.3615	0.6789	0.0994	0.9053	0.5155	0.5257	0.4021	0.7434	0.3925	0.6963
I-Mem	**0.4870***	**0.6457***	**0.3279***	**0.7755***	**0.0941**	**0.9508***	**0.4278***	**0.6386***	**0.3724***	**0.7757***	**0.3554***	**0.7228***

Table 1: The AUC performance of various models on the FIIF and PIIF setting. Best results are in bold; best baseline is underlined. * denotes p-value < 0.05 compared to the best-performing baseline. FIIF uses scalar α to control spurious feature–label correlation (larger α = stronger spurious correlation). PIIF uses (α, β) where α and β indicate the correlation of invariant and spurious features with the label, respectively.

Model	α=0.3		α=0.6		α=0.9		(α,β)=(0.7,0.8)		(α,β)=(0.9,0.8)		(α,β)=(0.9,0.7)	
	S ↓	C ↑	S ↓	C ↑	S ↓	C ↑	S ↓	C ↑	S ↓	C ↑	S ↓	C ↑
ERM	0.0369	0.1921	0.1288	0.0792	0.0673	0.0027	0.0829	0.0097	0.1324	0.0164	0.2468	0.0164
GroupDRO	0.0524	0.1609	0.1817	0.0657	0.0563	0.0022	0.0904	0.0102	0.1603	0.0095	0.1507	0.0097
IRM	0.0435	0.0795	0.1263	0.0425	0.0819	0.0014	0.0714	0.0012	0.1596	0.0027	0.1577	0.0038
IB-IRM	0.0959	0.0368	0.1572	0.0837	0.4092	0.0023	0.2361	0.0062	0.2582	0.0035	0.1459	0.0032
EIIL	0.0811	0.0650	0.1078	0.0632	0.0873	0.0021	0.0824	0.0020	0.1356	0.0112	0.1703	0.0067
TKNet	0.0438	0.1129	**0.1038**	0.1277	0.0565	0.0036	0.0776	0.0041	0.1499	0.0166	0.1701	0.0110
FOIL	0.0421	0.0314	0.1404	0.0704	0.0808	0.0032	0.0846	0.0002	0.1494	0.0122	0.2299	0.0041
I-Mem	**0.0342**	**0.2341**	0.1072	**0.2975**	**0.0559**	**0.0043**	**0.0702**	**0.0107**	**0.1294**	**0.0177**	**0.1361**	**0.0174**

Table 2: Feature importance comparison on FIIF and PIIF settings.The best results are shown in bold; the best among baseline methods is underlined. S is the average importance of spurious features, C of invariant features. A higher C indicates greater model attention to invariant factors, while a lower S reflects reduced reliance on spurious features.

Real-World Datasets. To assess generalization under diverse real-world conditions, we evaluated the model on three large-scale datasets: Anki, MaiMemo, and Duolingo (En2Fr and En2Es). These datasets differ in data distribution, user demographics (Chinese vs. Western learners), memory content (vocabulary vs. flashcards), and task settings (structured vs. self-paced learning). Such heterogeneity naturally induces distribution shifts, providing a strong testbed for out-of-distribution robustness. Detailed data set descriptions and heterogeneity analysis are provided in the supplementary material.

Baselines. To assess the effectiveness of our framework, we compare it against both classical memory models and OOD generalization methods. Classical baselines include HLR(Settles and Meeder 2016), Wickelgren(Wickelgren 1974), ACT-R(Anderson et al. 2004), and DAS3H(Choffin et al. 2019). OOD Generalization in Time Series include Group-DRO(Sagawa et al. 2019), IRM(Arjovsky et al. 2019), IB-IRM(Ahuja et al. 2021), EIIL(Creager, Jacobsen, and Zemel 2021), TKNET(Zeng et al. 2024), and FOIL(Liu et al. 2024a). All OOD methods, including ours, are built upon the same DKT(Piech et al. 2015) backbone to ensure fair comparison. Detailed descriptions of the baselines is provided in supplementary material.

Training Protocol. Simulation experiments are conducted under FIIF and PIIF settings, while real-world generalization is assessed via Leave-One-Domain-Out validation. MAE and AUC are used as evaluation metrics. All models are trained with Adam Optimizer(Kingma 2014), using an 80/20 train-test split and five independent runs for reliability. Details are provided in supplementary material.

Simulation Experiments

Performance Analysis (RQ1). The results on two synthetic datasets are presented in Table 1. I-Mem consistently achieves superior performance under varying degrees of spurious correlation. On the FIIF setting, I-Mem maintains optimal performance across different levels of spurious correlation and consistently outperforms all baselines. Notably, when the correlation between the spurious feature S and the target is 0.3 or 0.6, our model retains robust performance. In the PIIF setting, especially when the conditional entropy satisfies $H(S \mid Y) < H(C \mid Y)$, such as in configurations $(0.9, 0.8)$ and $(0.9, 0.7)$, environment-enhanced methods (e.g., GroupDRO, IRM) and environment-inference methods (e.g., EIIL) perform poorly. As the correlation between S and Y intensifies, these methods inevitably encode information from C through S, failing to correctly disentangle them. Furthermore, when $H(S \mid Y) > H(C \mid Y)$, all baseline methods exhibit catastrophic failures, indicating that it becomes extremely difficult to separate S from C, which impairs invariant learning. It is worth noting that FOIL does not perform well in identifying the spurious component C under these conditions, whereas I-Mem successfully isolates S through self-supervised global-local InfoMax, independent of the label Y. As a result, even under strong S-Y correla-

Model	En2Fr		En2Es		MaiMemo		Anki		Avg	
	MAE ↓	AUC ↑	MAE ↓	AUC ↑	MAE ↓	AUC ↑	MAE ↓	AUC ↑	MAE ↓	AUC ↑
HLR	0.5028	0.6206	0.4986	0.5797	0.6769	0.5012	0.6477	0.6504	0.5815	0.5880
Wickelgren	0.5163	0.5560	0.5138	0.5328	0.7069	0.5152	0.5347	0.6698	0.5679	0.5685
ACT-R	0.5034	0.5077	0.5012	0.5129	0.6741	0.5327	0.3627	0.5053	0.5104	0.5147
DAS3H	0.5516	0.4979	0.5353	0.4718	0.6932	0.5195	0.5583	0.6758	0.5846	0.5413
ERM	0.4997	0.6581	0.4255	0.6661	0.6795	0.5014	0.4475	0.6524	0.5343	0.6069
GroupDRO	0.5291	0.6198	0.4290	0.6769	0.6940	0.4651	0.3750	0.6199	0.5168	0.5954
IRM	0.6183	0.5501	0.6461	0.6639	0.6472	0.5206	0.5689	0.5423	0.6201	0.5942
IB-IRM	0.6169	0.5299	0.6606	0.4953	0.6715	0.5439	0.3776	0.5251	0.5817	0.5235
EIIL	0.6041	0.4821	0.6027	0.5197	0.6487	0.5537	0.4620	0.6875	0.5794	0.5608
TKNet	0.5607	0.5144	0.4044	0.5038	0.6939	0.5075	0.7846	0.5151	0.6609	0.5102
FOIL	0.5961	0.5938	0.4758	0.6666	0.6728	0.4992	0.3740	0.6852	0.5547	0.6112
I-Mem	**0.3608***	**0.6691***	**0.3829***	**0.6918***	**0.6159***	**0.5726***	**0.3654**	**0.7037***	**0.4312**	**0.6593**

Table 3: Generalization performance on real-world datasets. The best results are shown in bold; the best among baseline methods is underlined. * denotes p-value < 0.05 compared to the best-performing baseline.

tion, I-Mem shows an improved ability to reduce spurious interference and to better identify invariant components.

Feature Importance Validation (RQ3). To evaluate the effectiveness of I-Mem in distinguishing invariant features while suppressing spurious correlations, we conduct a systematic feature importance analysis across synthetic settings (see Table 2). We adopt a permutation-based AUC degradation metric (Gómez-Ramírez, Ávila-Villanueva, and Fernández-Blázquez 2020) to quantify model reliance on spurious (S) and invariant (C) factors under varying correlation . The results demonstrate that I-Mem consistently exhibits the lowest reliance on spurious features and the strongest attention to invariant ones across all configurations, significantly outperforming baselines such as ERM, IRM, GroupDRO, and FOIL. Specifically, under high spurious correlation conditions ($\alpha = 0.6$), I-Mem reduces dependency on spurious features by 16.9% compared to ERM, and increases attention to invariant features by 275.1%. This consistent superiority across both FIIF and PIIF scenarios supports I-Mem's robustness at the feature level and highlights its capability to model stable cognitive structures, aligning with the core objectives of invariant representation learning.

Real-World Dataset Evaluation

Performance Verification (RQ1). The effectiveness of I-Mem is further validated on real-world memory time-series datasets. As shown in Table 3, I-Mem consistently outperforms all competing methods, achieving state-of-the-art performance across all datasets. Compared with environment-inference (EIIL and FOIL) and environment-enhancement (IRM) algorithms, I-Mem demonstrates significant improvements, indicating its superior ability to identify spurious features S via InfoMax-based self-supervised learning, which in turn facilitates the acquisition of invariant representations. It is noteworthy that traditional memory equations generalize well on the Anki dataset but perform poorly on En2Fr. This highlights the limitation of global descriptors in capturing the fine-grained memory dynamics required in language learning tasks, thereby reducing model adaptability. FOIL infers latent environments via feature-space clustering, but this approach may struggle to capture true invariant signals under noise, potentially limiting representation quality. TKNet models temporal evolution across domains using fixed-order transitions, which may limit its flexibility in capturing local shifts and learner-specific variability common in real-world memory data. This constraint can affect its ability to generalize invariant patterns under heterogeneous conditions.

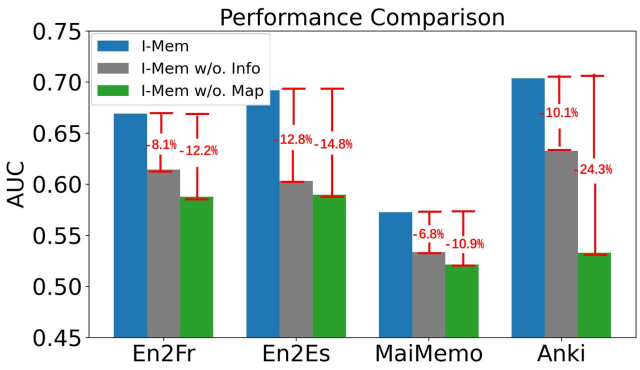

Figure 3: Evaluating the impact of removing InfoMax and Mapping modules on AUC

$\mathcal{H}$-Num	En2Fr	En2Es	MaiMemo	Anki
$\mathcal{H}$-1	0.6691	**0.6918**	0.5726	0.7037
$\mathcal{H}$-2	**0.6724**	0.6872	**0.5969**	0.7058
$\mathcal{H}$-3	0.6675	0.6814	0.5802	**0.7139**
$\mathcal{H}$-4	0.6676	0.6869	0.5654	0.7125
$\mathcal{H}$-5	0.6691	0.6812	0.5692	0.7056

Table 4: Effect of Spurious Set ($\mathcal{H}$) Size on Generalization.

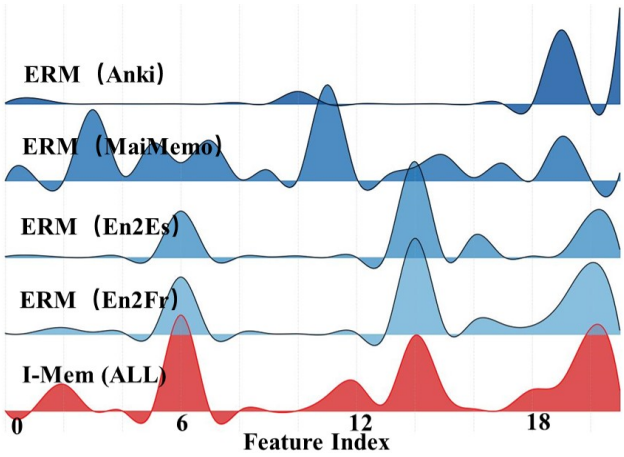

Figure 4: Feature importance distributions.

Index	Name	Importance
6	Recall Decay Score	0.0416
20	Review Decay Indicator	0.0358
14	Incorrect in Last 3	0.0328
19	Overall User Accuracy	0.0126
12	Accuracy in Last 3	0.0126
2	Days Since First Review	0.0116
18	Wrong Streak Length	0.0093
21	Previous Response Outcome	0.0090

Table 5: Feature Importance Ranking

Ablation Studies (RQ2)

Module Ablation. To assess the contribution of each component in our framework, we perform ablation experiments using three variants: (1) I-Mem w/o. Info: The InfoMax step is removed, and the encoder is replaced with a random module. (2) I-Mem w/o. Map: The mapping step is removed, and high-dimensional representations are directly used as the decorrelation environment S.As shown in Table 3, removing either InfoMax or the mapping step leads to performance degradation across all four datasets in terms of AUC. In particular, on the Anki dataset, removing InfoMax results in a 10% drop in AUC, while removing the mapping step leads to a 24% decrease. These results indicate that the InfoMax mechanism helps extract stable and discriminative representations, while the mapping step plays a critical role in suppressing high-dimensional noise and improves decorrelation effectiveness. Overall, their synergy is essential for enhancing the robustness and generalization of the model.

Size of the Self-Supervised Spurious Set $\mathcal{H}$. We also conduct an ablation study on the size of the latent spurious set $\mathcal{H}$. Specifically, we vary the number of spurious samples to observe its effect on model performance. As shown in Table 4, the model achieves optimal AUC performance when the size of the spurious set is moderate. In this case, the model can clearly distinguish between true and spurious features, leading to more stable and discriminative representations. However, when the spurious set is too large, the model tends to overly focus on avoiding spurious patterns during training, thereby neglecting the fine-grained modeling of real data. This training shift ultimately results in decreased AUC performance on real-world datasets.

Application (RQ4)

To validate the utility of I-Mem in real-world memory behavior modeling, we conducted a cross-environment feature attribution analysis, focusing on its ability to identify robust, generalizable features while suppressing environment-specific spurious signals. As shown in Figure 4, I-Mem exhibits more stable attribution patterns across environments. Notably, Feature 6 ("Recall Decay Score") consistently ranks highest, reflecting its role in modeling decay as described in the Ebbinghaus forgetting curve. In contrast, models trained with ERM display considerable variability in feature dependence across environments—for example, ERM (Anki) emphasizes Features 19 and 21, while ERM (MaiMemo) prioritizes different subsets—indicating a tendency to overfit patterns.

Further analysis, as shown in Table 5, demonstrates that I-Mem consistently amplifies theoretically grounded features such as Feature 6 ("Recall Decay Score") and Feature 19 ("Overall User Accuracy"). These features correspond to well-established cognitive mechanisms, including time-based decay, long-term proficiency estimation, and sequential response dependency, aligning with activation strength and memory trace accumulation in ACT-R and related frameworks. Beyond theory-aligned signals, I-Mem also reliably highlights short-term performance indicators such as Feature 14 ("Incorrect in Last 3") and Feature 12 ("Accuracy in Last 3"). While not explicitly modeled in classical cognitive theories, these features may capture transient recall variability or context-sensitive learning effects, suggesting that I-Mem is capable of uncovering latent cognitive patterns beyond existing frameworks. This pattern is further supported by supplementary materials, which show that these features significantly contribute to cross-domain generalization and play a critical role in controlled feature selection experiments. Overall, the alignment of I-Mem's attributions with both established and novel cognitive signals underscores its interpretability and potential to reveal generalizable behavioral patterns grounded in memory theory.

Conclusion

This study addresses the challenge of out-of-distribution generalization in memory behavior modeling by proposing I-Mem, a self-supervised invariant representation learning framework that does not require environment labels. Comprehensive experiments on both synthetic and real-world datasets demonstrate that I-Mem can effectively reduce spurious correlations and highlight relatively stable behavioral patterns. Feature attribution analysis reveals that the invariant features extracted by the model are broadly consistent with classical memory effects, such as recall decay. Moreover, I-Mem also identifies several short-term behavioral patterns that may reflect latent cognitive dynamics beyond existing theories, offering potential directions for future research.

Acknowledgements

This work was joint supported by the National Science and Technology Major Project (2022ZD0117103), National Natural Science Foundation of China (62293554, 62577028). Youth Al Talents Fund of the Chinese Association of Automation under Major Program(HBRC-JKYZD-2024-310), Higher Education Science Research Program of China Association of Higher Education (23XXK0301), Hubei Provincial Natural Science Foundation of China (2023AFA020), Fundamental Research Funds for the Central Universities (CCNU25AI005), and China Postdoctoral Science Foundation (2024M761088).

References

Ahuja, K.; Caballero, E.; Zhang, D.; Gagnon-Audet, J.-C.; Bengio, Y.; Mitliagkas, I.; and Rish, I. 2021. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34: 3438–3450.

Anderson, J. R.; Bothell, D.; Byrne, M. D.; Douglass, S.; Lebiere, C.; and Qin, Y. 2004. An integrated theory of the mind. *Psychological review*, 111(4): 1036.

Anderson, J. R.; and Schooler, L. J. 1991. Reflections of the environment in memory. *Psychological science*, 2(6): 396–408.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bayat, R.; Pezeshki, M.; Dohmatob, E.; Lopez-Paz, D.; and Vincent, P. 2024. The Pitfalls of Memorization: When Memorization Hurts Generalization. *arXiv preprint arXiv:2412.07684*.

Carvalho, W.; and Lampinen, A. 2025. Naturalistic Computational Cognitive Science: Towards generalizable models and theories that capture the full range of natural behavior. *arXiv preprint arXiv:2502.20349*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Chen, Y.; Bian, Y.; Zhou, K.; Xie, B.; Han, B.; and Cheng, J. 2023. Does invariant graph learning via environment augmentation learn invariance? *Advances in Neural Information Processing Systems*, 36: 71486–71519.

Choffin, B.; Popineau, F.; Bourda, Y.; and Vie, J.-J. 2019. DAS3H: modeling student learning and forgetting for optimally scheduling distributed practice of skills. *arXiv preprint arXiv:1905.06873*.

Creager, E.; Jacobsen, J.-H.; and Zemel, R. 2021. Environment inference for invariant learning. In *International Conference on Machine Learning*, 2189–2200. PMLR.

Ebbinghaus. 2013. Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4): 155.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gómez-Ramírez, J.; Ávila-Villanueva, M.; and Fernández-Blázquez, M. Á. 2020. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Scientific reports*, 10(1): 20630.

Goyal, A.; and Bengio, Y. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266): 20210068.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, H.; Kamarthi, H.; Kong, L.; Zhao, Z.; Zhang, C.; and Prakash, B. A. 2024a. Time-series forecasting for out-of-distribution generalization using invariant learning. *arXiv preprint arXiv:2406.09130*.

Liu, H.; Kamarthi, H.; Kong, L.; Zhao, Z.; Zhang, C.; and Prakash, B. A. 2024b. Time-series forecasting for out-of-distribution generalization using invariant learning. *arXiv preprint arXiv:2406.09130*.

Liu, J.; Shen, Z.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Liu, S.; Li, Q.; Shen, X.; Sun, J.; and Yang, Z. 2024c. Automated discovery of symbolic laws governing skill acquisition from naturally occurring data. *Nature Computational Science*, 4(5): 334–345.

Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; and Luo, W. 2023. simpleKT: a simple but tough-to-beat baseline for knowledge tracing. *arXiv preprint arXiv:2302.06881*.

Lu, C.; Wu, Y.; Hernández-Lobato, J. M.; and Schölkopf, B. 2021. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*.

Ma, B.; Hettiarachchi, G. P.; Fukui, S.; and Ando, Y. 2023. Each encounter counts: Modeling language learning and forgetting. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, 79–88.

Mozer, M. C.; and Lindsey, R. V. 2016. Predicting and improving memory retention: Psychological theory matters in the big data era. In *Big data in cognitive science*, 43–73. Psychology Press.

Pearlin, E.; and Gandhi, S. M. G. 2024. Enhancing User Behavior Analysis in Mobile Language Learning Apps Through Gamification and AI Integration: A Transformer-Based Deep Learning Approach. In *2024 International Conference on Data Science and Network Security (ICDSNS)*, 1–6. IEEE.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal

representation learning. *Proceedings of the IEEE*, 109(5): 612–634.

Settles, B.; and Meeder, B. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, 1848–1858.

Shen, X.; Hu, Z.; Chen, D.; Sun, J.; and Liu, S. 2025. ROKAN: Toward Interpretable and Domain-Robust Memory Behavior Modeling. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, 2588–2598. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720406.

Shen, X.; Hu, Z.; Chen, Q.; and Wang, P. 2026. Evolvable psychology informed neural network for memory behavior modeling. *Information Processing & Management*, 63(1): 104312.

Sun, J.; Chen, Q.; Huang, Z.; Hu, Z.; Liang, R.; and Shen, X. 2025. Combining Denoised Neural Network and Genetic Symbolic Regression for Memory Behavior Modeling via Dynamic Asynchronous Optimization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2735–2746.

Vapnik, V. N. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5): 988–999.

Walkington, C. A. 2013. Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of educational psychology*, 105(4): 932.

Walsh, M. M.; Gluck, K. A.; Gunzelmann, G.; Jastrzembski, T.; and Krusmark, M. 2018. Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive science*, 42: 644–691.

Wickelgren, W. A. 1974. Single-trace fragility theory of memory dynamics. *Memory & Cognition*, 2(4): 775–780.

Wixted, J. T. 2004. The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.*, 55(1): 235–269.

Wu, X.; Teng, F.; Li, X.; Zhang, J.; Li, T.; and Duan, Q. 2025. Out-of-Distribution Generalization in Time Series: A Survey. *arXiv preprint arXiv:2503.13868*.

Wu, Y.-X.; Wang, X.; Zhang, A.; He, X.; and Chua, T.-S. 2022. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872*.

Ye, J.; Su, J.; and Cao, Y. 2022. A stochastic shortest path algorithm for optimizing spaced repetition scheduling. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 4381–4390.

Ye, W.; Zheng, G.; Cao, X.; Ma, Y.; and Zhang, A. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Zeng, Q.; Wang, W.; Zhou, F.; Xu, G.; Pu, R.; Shui, C.; Gagné, C.; Yang, S.; Ling, C. X.; and Wang, B. 2024. Generalizing across temporal domains with koopman operators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16651–16659.