Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ContentMIRT: Context-Augmented Multidimensional Item Response Theory with Neural Latent Regression and Sparse Moderation

Yong Zhang

hflz_zy@126.com

Central China Normal University

Bo Li

Central China Normal University

Hong Qin

Zhongnan University of Economics and Law

Article

Keywords: Multidimensional Item Response Theory, Variational Autoencoders, Latent Regression, Differential Item Functioning, Process Paradata, Measurement Invariance

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10972910/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

ContentMIRT: Context-Augmented Multidimensional Item Response Theory with Neural Latent Regression and Sparse Moderation

Yong Zhang[1*], Bo Li[1] and Hong Qin[2]

[1*]School of Mathematics and Statistics, Central China Normal University, Luoyu Road, Wuhan, 430079, HuBei, China.
[2]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Nanhu Avenue, Wuhan, 430073, HuBei, China.

*Corresponding author(s). E-mail(s): hflz_zy@126.com;

Abstract

In large-scale educational assessments (e.g., TIMSS, PISA), item calibration and background conditioning are conventionally conducted via multi-stage pipelines, where linear latent regression fails to capture non-linear saturation thresholds while ignoring item-level Differential Item Functioning (DIF). To address these challenges, we propose ContentMIRT (C-MIRT), an end-to-end variational deep generative measurement framework that unifies multidimensional item response modeling, non-linear neural latent regression, and sparse moderation. C-MIRT incorporates a Tabular Neural Context Prior to map heterogeneous background variables and digital process timing paradata directly into individualized latent trait prior distributions. Simultaneously, an $\boldsymbol{L_1}$-regularized Sparse Moderation Decoder isolates item-specific DIF shifts, decoupling construct proficiency from measurement bias under formal majority-invariance conditions. To ensure metric determinacy without ad-hoc rotational restrictions, we employ positive softplus discrimination constraints alongside an exact Orthogonal Procrustes alignment protocol. Across Monte Carlo simulations spanning 30% to 85% response sparsity, C-MIRT consistently surpasses standard and linear latent regression MIRT in out-of-item predictive accuracy and discrimination recovery, while two-stage screening secures high DIF detection precision. Across six empirical assessment cohorts from TIMSS 2023 and PISA 2022 spanning three performance tiers (Singapore, USA, Chile), C-MIRT exhibits strong linear alignment with operational Plausible Values ($\boldsymbol{r_\theta > 0.81}$), yields a 26.0% (PISA Singapore) to 39.2% (TIMSS Singapore) reduction in posterior trait uncertainty, and reveals via gradient attributions that behavioral timing paradata account for over 55%

of prior predictive influence. These findings demonstrate that integrating neural latent regression and sparse moderation provides an expressive, interpretable, and scalable framework for context-augmented cognitive measurement.

Keywords: Multidimensional Item Response Theory, Variational Autoencoders, Latent Regression, Differential Item Functioning, Process Paradata, Measurement Invariance

1 Introduction

In large-scale educational and psychological assessments, Multidimensional Item Response Theory (MIRT; Lord 1980; Embretson and Reise 2000; Reckase 2009) models the joint probability of observed categorical responses as a function of continuous latent competencies and item parameters, including multidimensional discrimination, difficulty thresholds, and pseudo-guessing. Traditional calibration algorithms—such as Marginal Maximum Likelihood via Expectation-Maximization (Bock and Aitkin 1981), Markov Chain Monte Carlo (Patz and Junker 1999), and Metropolis-Hastings Robbins-Monro algorithms (Cai 2010; Chalmers 2012)—routinely specify an unconditional standard normal prior $\boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for latent abilities.

However, modern computer-based assessments routinely capture rich auxiliary student background information, including socio-demographic indicators, disciplinary self-efficacy measures, co-administered cognitive tasks, and digital process timing logs. Classical MIRT isolates item calibration from these auxiliary variables, separating background conditioning into a secondary, multi-stage imputation step, such as operational Plausible Value generation (von Davier and Sinharay 2010). While multi-stage conditioning prevents item parameter drift during operational reporting, it introduces substantial methodological limitations: linear latent regression (Adams et al. 1997) cannot accommodate non-linear returns or interactive thresholds across background predictors (De Boeck and Wilson 2004), while unregularized conditioning models risk conflating population-level proficiency differences with item-specific measurement biases, known as Differential Item Functioning (DIF; Holland and Wainer 1993).

To unify item calibration and background conditioning within a coherent generative framework, we propose ContentMIRT (C-MIRT), a variational deep generative measurement model that jointly models cognitive response data, non-linear neural latent regression, and sparse item-level moderation. This work makes four principal methodological contributions:

1. First, we formulate a Tabular Neural Context Prior, $p(\boldsymbol{\theta} \mid \mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}_0(\mathbf{X}), \text{diag}(\boldsymbol{\sigma}_0^2(\mathbf{X})))$, parameterized by a deep neural network with hyperbolic tangent activations to capture non-linear covariate saturation and multi-way socio-cognitive interactions.
2. Second, we introduce a Sparse Moderation Decoder that incorporates an L_1-regularized difficulty moderation term $\boldsymbol{\gamma}_j^T \mathbf{X}_i$ into the response probability layer, decoupling construct variance from item-level Differential Item Functioning under majority-invariance regularity conditions.

3. Third, we establish theoretical identifiability constraints through positive softplus parameterization and analytical Kullback-Leibler divergence, accompanied by a post-hoc Orthogonal Procrustes Alignment protocol (Schönemann 1966) to evaluate split-half measurement invariance under rotational indeterminacy.
4. Fourth, we conduct extensive 5-seed empirical evaluations across 6 assessment cohorts from TIMSS 2023 and PISA 2022 spanning three distinct educational performance tiers (Singapore, United States, and Chile), combined with granular item-level and student-level diagnostic analyses connecting survey questionnaire variables directly to psychometric parameters.

The remainder of this article is organized as follows. Section 2 reviews related literature on latent regression, process timing paradata, regularized DIF estimation, and deep IRT. Section 3 formalizes the mathematical architecture, identifiability properties, loss functions, and optimization algorithms of C-MIRT. Section 4 describes the Monte Carlo simulation and empirical assessment designs. Section 5 analyzes the quantitative experimental results. Section 6 presents granular item-level and student-level diagnostic case studies. Section 7 provides a discussion of methodological implications, empirical trade-offs, limitations, and future directions. Appendix A details diagnostic and sensitivity analyses, including ELBO convergence, local item independence, prior hyperparameter sensitivity, robustness against alternative non-linear data-generating processes, and a systematic factorial ablation study.

2 Literature Review and Theoretical Positioning

2.1 Latent Regression and Conditioning Models in Psychometrics

The integration of auxiliary background covariates into latent trait estimation was introduced by Mislevy (1991) and formalized in multilevel generalized linear frameworks by Adams et al. (1997). Under classical Latent Regression MIRT, examinee abilities are linearly regressed onto background predictors:

$$\boldsymbol{\theta}_i = \boldsymbol{\Gamma}_{\text{LR}}^T \mathbf{X}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{1}$$

This formulation underpins Plausible Value methodology in TIMSS, PIRLS, and PISA (von Davier and Sinharay 2010). However, linear latent regression assumes constant slope coefficients across the entire covariate range. In educational settings, variables such as study duration, parental involvement, and academic anxiety frequently exhibit non-linear thresholds or ceiling effects (De Boeck and Wilson 2004). C-MIRT extends latent regression by parameterizing conditional prior distributions through deep neural encoders, capturing non-linear functional relationships while obviating the need for pre-specified polynomial or interaction bases.

2.2 Regularized Detection of Differential Item Functioning

Differential Item Functioning occurs when examinees from different background subgroups with identical latent ability levels have different probabilities of endorsing an

item (Holland and Wainer 1993; Millsap 2011; Magis et al. 2010). In high-dimensional contexts with multiple continuous and categorical background variables, classical pairwise contingency table methods (e.g., Mantel-Haenszel) become cumbersome. To address this, penalized maximum likelihood approaches have been developed for Rasch and IRT models (Tutz and Schauberger 2015), as well as regularized moderated non-linear factor analysis (Bauer et al. 2020; Belzak and Bauer 2020). These methods demonstrate that applying Lasso-type penalties (Tibshirani 1996) to item-by-covariate interaction terms enables automated DIF screening while shrinking non-DIF parameters to zero. C-MIRT embeds this regularized moderation principle directly into a deep variational measurement decoder, allowing joint estimation of non-linear priors and sparse item moderation.

2.3 Process Timing Paradata and Behavioral Engagement in Measurement

Computer-based assessment platforms provide detailed timing logs and audit trails. van der Linden (2007) formalized the hierarchical framework modeling response accuracy and response speed as distinct, correlated latent traits. Although mathematically principled, joint dual-construct models double the latent dimensionality and typically require computationally intensive Markov Chain Monte Carlo sampling. In C-MIRT, summary indices of cognitive processing speed (e.g., student mean response time) and metacognitive pacing consistency (e.g., standard deviation of response times) are incorporated directly as continuous conditioning covariates within the neural context prior. This feature-conditioning strategy enables the model to incorporate behavioral engagement patterns to inform latent ability priors without increasing the dimensionality of the latent space.

2.4 Deep Variational Inference for Item Response Theory

Recent psychometric research has applied Variational Autoencoders (VAEs; Kingma and Welling 2014; Blei et al. 2017) to scale IRT calibration to massive examinee cohorts (Wu et al. 2020; Urban and Bauer 2021; Cúri et al. 2019; Tsutsumi et al. 2022). Variational IRT models employ an inference network (encoder) to approximate intractable posterior distributions $q(\boldsymbol{\theta} \mid \mathbf{Y})$ while optimizing an Evidence Lower Bound (ELBO). While scalable extensions such as the Importance-Weighted Amortized Variational Estimator (I-WAVE; Urban and Bauer 2021) support latent regression by regressing factor means onto background covariates, these formulations remain inherently linear, do not accommodate multi-way non-linear feature interactions or saturation thresholds, and lack joint mechanisms for sparse item-level DIF moderation and process timing paradata integration.

Furthermore, existing deep IRT formulations present two fundamental challenges. First, unconstrained neural encoders yield arbitrary coordinate rotations and sign flips across independent calibration runs, complicating metric comparisons across samples or test forms. Second, standard VAE-IRT models route all background information through the low-dimensional latent bottleneck $\boldsymbol{\theta}$, which can conflate group-level ability differences with item-specific Differential Item Functioning. C-MIRT addresses both

issues through positive softplus discrimination constraints, closed-form Gaussian KL divergence against dynamic neural priors, and an L_1-regularized Sparse Moderation Decoder.

3 Methodology: The ContentMIRT Architecture

3.1 Mathematical Notation

The Notation Table summarizes the mathematical symbols, tensor shapes, and operational definitions utilized throughout this paper.

Notation Table: *Summary of Mathematical Notations and Tensor Dimensions*

Symbol	Dimensionality	Description
N, J, D, d_{bg}	$\mathbb{N}$	Sample size, item pool size, latent trait dimension, background feature dimension
$\mathbf{Y}_i$	$\{0,1\}^J$	Observed binary response vector for examinee i
$\mathbf{M}_i$	$\{0,1\}^J$	Administration mask ($M_{ij}=1$ if item j was administered to examinee i)
$\mathbf{Z}_i$	$\mathbb{R}^{2J}$	Masked input vector $\mathbf{Z}_i = [\mathbf{Y}_i \odot \mathbf{M}_i \,\|\, \mathbf{M}_i]$ for the posterior encoder
$\mathbf{X}_i$	$\mathbb{R}^{d_{\text{bg}}}$	Background covariate and process paradata vector for examinee i
$\boldsymbol{\theta}_i$	$\mathbb{R}^D$	Multidimensional continuous latent ability vector for examinee i
$\boldsymbol{\mu}_0(\mathbf{X}_i), \boldsymbol{\sigma}_0^2(\mathbf{X}_i)$	$\mathbb{R}^D$	Mean and variance vectors of the Tabular Neural Context Prior $p(\boldsymbol{\theta}_i \mid \mathbf{X}_i)$
$\boldsymbol{\mu}_q(\mathbf{Z}_i), \boldsymbol{\sigma}_q^2(\mathbf{Z}_i)$	$\mathbb{R}^D$	Mean and variance vectors of the Variational Posterior $q(\boldsymbol{\theta}_i \mid \mathbf{Z}_i)$
$\boldsymbol{a}_j$	$\mathbb{R}_+^D$	Positively constrained multidimensional discrimination vector for item j
$\mathbf{A}$	$\mathbb{R}^{J\times D}$	Full test discrimination matrix $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_J]^T$
b_j	$\mathbb{R}$	Scalar difficulty threshold parameter for item j (intercept $d_j = -b_j$)
$\boldsymbol{\gamma}_j$	$\mathbb{R}^{d_{\text{bg}}}$	Sparse background moderation (DIF) coefficient vector for item j
$\boldsymbol{\Gamma}$	$\mathbb{R}^{J\times d_{\text{bg}}}$	Full test sparse DIF moderation matrix $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J]^T$
$\mathbf{R}$	$\mathbb{R}^{D\times D}$	Orthogonal Procrustes rotation matrix ($\mathbf{R}^T\mathbf{R} = \mathbf{I}$)
$r(\mathbf{A}_{rot})$	$[-1,1]$	Post-Procrustes discrimination correlation $r(\mathbf{A}_A, \mathbf{A}_B\mathbf{R})$
$\bar{r}_\theta^{(proc)}$	$[-1,1]$	Procrustes-aligned mean latent ability correlation in simulation
$r_\theta^{(max)}$	$[0,1]$	Held-out maximum latent trait correlation with operational Plausible Values
$I(x_k)$	$\mathbb{R}_+$	Relative gradient-based feature attribution percentage for covariate x_k
$\beta, \lambda_{\text{dif}}$	$\mathbb{R}_+$	Hyperparameter weights for KL divergence and L_1 moderation regularization

3.2 Generative Measurement Model & Sparse Moderation Decoder

Let $Y_{ij} \in \{0,1\}$ denote the dichotomous response of student $i \in \{1,\ldots,N\}$ to item $j \in \{1,\ldots,J\}$. Under the compensatory multidimensional logistic formulation augmented with sparse background moderation, the conditional probability of a correct response given latent ability $\boldsymbol{\theta}_i$ and background vector $\mathbf{X}_i$ is defined as:

$$P(Y_{ij}=1 \mid \boldsymbol{\theta}_i, \mathbf{X}_i) = \sigma\left(\boldsymbol{a}_j^T\boldsymbol{\theta}_i - b_j + \boldsymbol{\gamma}_j^T\mathbf{X}_i\right) \tag{2}$$

where $\sigma(z) = \frac{1}{1+\exp(-z)}$ denotes the standard logistic sigmoid function, and b_j denotes the scalar difficulty threshold (corresponding to intercept $d_j = -b_j$).

To ensure metric interpretability and prevent sign-flip indeterminacy, the discrimination vectors $\boldsymbol{a}_j$ are parameterized via a strictly positive softplus transformation

with a numerical offset:

$$\boldsymbol{a}_j = \log(1 + \exp(\boldsymbol{a}_{j,raw})) + 10^{-3} > \mathbf{0} \tag{3}$$

In standardized educational achievement assessments, test items are intentionally constructed to evaluate positive proficiencies monotonically; hence, the positive manifold constraint $\boldsymbol{a}_j > \mathbf{0}$ is substantively grounded (cf. van der Maas et al. 2006). However, we emphasize that this positivity constraint is methodologically tailored to monotonic cognitive ability and achievement domains. For non-cognitive instruments (e.g., personality inventories, clinical psychiatric scales, or attitudinal surveys) that feature bipolar constructs or reverse-coded items, this positivity restriction can be readily relaxed: item discriminations can be parameterized as unconstrained real values $\boldsymbol{a}_j \in \mathbb{R}^D$, while sign-reflection indeterminacy ($\boldsymbol{\theta}_d \to -\boldsymbol{\theta}_d$) can be resolved by pre-specifying the positive polarity of a minimal set of designated "sign anchor items" (e.g., fixing $a_{j^*,d} > 0$ for one prototypical item per dimension; cf. Chen et al. 2019). Furthermore, while the present study focuses on dichotomously scored items, the logistic measurement layer can be straightforwardly extended to polytomous items with ordered categorical responses via multidimensional graded response formulations (Samejima 1969).

The vector $\boldsymbol{\gamma}_j = [\gamma_{j,1}, \ldots, \gamma_{j,d_{\text{bg}}}]^T \in \mathbb{R}^{d_{\text{bg}}}$ represents the item-specific background moderation shift. When $\gamma_{j,k} \neq 0$, the effective difficulty of item j shifts as a function of covariate $x_{i,k}$, capturing Differential Item Functioning (DIF). By penalizing the moderation matrix $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J]^T$ with an L_1 penalty, the model shrinks benign moderation coefficients toward zero, allowing $\boldsymbol{\gamma}_j$ to activate selectively on items exhibiting measurement bias. From a measurement identifiability standpoint, separating population ability impact ($\boldsymbol{\mu}_0(\mathbf{X}_i)$) from item-specific measurement bias ($\boldsymbol{\gamma}_j^T \mathbf{X}_i$) is guaranteed under a formal majority-invariance condition (detailed in Proposition 1); in fully confirmatory settings, formal identifiability can also be established by designating known anchor items ($\boldsymbol{\gamma}_{anchor} = \mathbf{0}$) or imposing a sum-to-zero constraint ($\sum_{j=1}^{J} \boldsymbol{\gamma}_j = \mathbf{0}$). Methodologically, this formulation specifically identifies Uniform DIF (shifts in intercept/difficulty). Extensions to Non-uniform DIF (shifts in multidimensional discrimination) could be achieved by parameterizing $\boldsymbol{a}_j$ as a neural output of $\mathbf{X}_i$, though this remains beyond the scope of the current uniform-DIF Sparse Moderation Decoder.

3.3 Tabular Neural Context Prior

In contrast to standard MIRT which enforces an unconditional prior $\boldsymbol{\theta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, C-MIRT specifies an examinee-specific conditional prior distribution:

$$p(\boldsymbol{\theta}_i \mid \mathbf{X}_i) = \mathcal{N}\left(\boldsymbol{\mu}_0(\mathbf{X}_i), \operatorname{diag}(\boldsymbol{\sigma}_0^2(\mathbf{X}_i))\right) \tag{4}$$

The prior mean $\boldsymbol{\mu}_0(\mathbf{X}_i)$ and prior log-variance $\log \boldsymbol{\sigma}_0^2(\mathbf{X}_i)$ are parameterized by a deep Tabular Neural Context Encoder:

$$\mathbf{h}_1(\mathbf{X}_i) = \tanh\left(\mathbf{W}_1 \mathbf{X}_i + \mathbf{b}_1\right), \quad \mathbf{h}_2(\mathbf{X}_i) = \tanh\left(\mathbf{W}_2 \mathbf{h}_1(\mathbf{X}_i) + \mathbf{b}_2\right) \tag{5}$$

$$\boldsymbol{\mu}_0(\mathbf{X}_i) = \mathbf{W}_3\mathbf{h}_2(\mathbf{X}_i) + \mathbf{b}_3 \tag{6}$$

$$\log \boldsymbol{\sigma}_0^2(\mathbf{X}_i) = \text{clip}\left(\mathbf{W}_4\mathbf{h}_2(\mathbf{X}_i) + \mathbf{b}_4, -6.0, 4.0\right) \tag{7}$$

The Tanh activation functions naturally accommodate non-linear saturation thresholds and multi-way feature interactions across background questionnaire variables. Clamping the log-variance to $[-6.0, 4.0]$ ensures numerical stability during backpropagation. While the conditional prior covariance $\text{diag}(\boldsymbol{\sigma}_0^2(\mathbf{X}_i))$ specifies conditional independence given $\mathbf{X}_i$, the marginal latent distribution $p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta} \mid \mathbf{X})p(\mathbf{X})d\mathbf{X}$ naturally recovers correlated, oblique factor structures across dimensions because the shared background covariates $\mathbf{X}$ jointly drive the multi-output conditional expectations $\boldsymbol{\mu}_0(\mathbf{X})$. For applications requiring explicit conditional correlation modeling, full-rank Cholesky covariance factorizations $\boldsymbol{\Sigma}_0(\mathbf{X}_i) = \mathbf{L}_0(\mathbf{X}_i)\mathbf{L}_0(\mathbf{X}_i)^T$ can be directly parameterized by expanding the context encoder's output heads.

3.4 Variational Theta Posterior Encoder

Because the true posterior distribution $p(\boldsymbol{\theta}_i \mid \mathbf{Y}_i, \mathbf{M}_i, \mathbf{X}_i)$ involves an intractable integral, we employ an amortized variational approximate posterior:

$$q(\boldsymbol{\theta}_i \mid \mathbf{Z}_i) = \mathcal{N}\left(\boldsymbol{\mu}_q(\mathbf{Z}_i), \text{diag}(\boldsymbol{\sigma}_q^2(\mathbf{Z}_i))\right) \tag{8}$$

Background covariates $\mathbf{X}_i$ are excluded from the posterior encoder input. This architectural separation ensures that background information influences latent ability estimates exclusively through the KL-regularized prior channel (Equation 13), rather than bypassing the principled Bayesian shrinkage mechanism via a direct posterior pathway. If $\mathbf{X}_i$ were routed directly into $q(\boldsymbol{\theta}_i \mid \mathbf{Z}_i, \mathbf{X}_i)$, a high-capacity deep inference network could easily exploit direct "shortcut learning"—mapping demographic labels and background predictors straight to latent scores while attenuating attention to sparse cognitive response patterns $\mathbf{Y}_i$. Such direct posterior conditioning would not only violate fundamental psychometric measurement ethics (wherein proficiency reporting must remain strictly anchored to demonstrated item response evidence; Dorans and Cook 2018; Fischer et al. 2023), but would also disable the analytical Bayesian shrinkage mechanism: background signals would leak directly into point estimates without being appropriately moderated by the prior's conditional uncertainty $\boldsymbol{\sigma}_0^2(\mathbf{X}_i)$. Structural isolation thus preserves an interpretable, principled decomposition between empirical cognitive evidence and auxiliary collateral information.

To handle matrix-sampled assessment designs without imputing arbitrary values for unadministered items, the encoder receives the $2J$-dimensional concatenated tensor $\mathbf{Z}_i = [\mathbf{Y}_i \odot \mathbf{M}_i \,\|\, \mathbf{M}_i] \in \mathbb{R}^{2J}$, explicitly separating missing items ($M_{ij} = 0$) from incorrect responses ($Y_{ij} = 0, M_{ij} = 1$):

$$\mathbf{e}_1(\mathbf{Z}_i) = \text{ReLU}\left(\mathbf{W}_{q,1}\mathbf{Z}_i + \mathbf{b}_{q,1}\right), \quad \mathbf{e}_2(\mathbf{Z}_i) = \text{ReLU}\left(\mathbf{W}_{q,2}\mathbf{e}_1(\mathbf{Z}_i) + \mathbf{b}_{q,2}\right) \tag{9}$$

$$\boldsymbol{\mu}_q(\mathbf{Z}_i) = \mathbf{W}_{q,3}\mathbf{e}_2(\mathbf{Z}_i) + \mathbf{b}_{q,3}, \quad \log \boldsymbol{\sigma}_q^2(\mathbf{Z}_i) = \text{clip}\left(\mathbf{W}_{q,4}\mathbf{e}_2(\mathbf{Z}_i) + \mathbf{b}_{q,4}, -6.0, 4.0\right) \tag{10}$$

During training, latent ability vectors $\boldsymbol{\theta}_i$ are sampled via Gaussian reparameterization:

$$\boldsymbol{\theta}_i = \boldsymbol{\mu}_q(\mathbf{Z}_i) + \boldsymbol{\sigma}_q(\mathbf{Z}_i) \odot \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{11}$$

During deterministic ability scoring (with stochastic variational sampling disabled), the posterior expectation $\hat{\boldsymbol{\theta}}_i = \boldsymbol{\mu}_q(\mathbf{Z}_i)$ is returned directly, ensuring deterministic point scoring.

The overall neural architecture is illustrated in Figure 1, and the end-to-end calibration procedure is summarized in Algorithm 1.

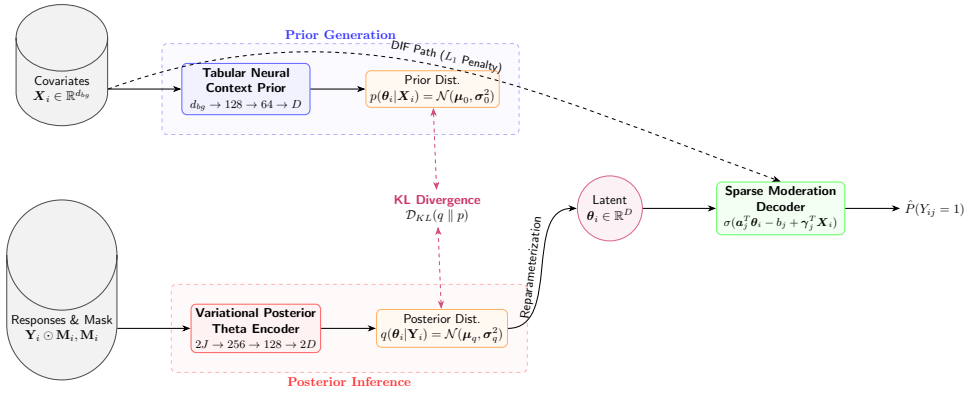


Fig. 1 ContentMIRT architecture overview. The framework couples a Tabular Neural Context Prior (mapping background covariates $\mathbf{X}_i$ to conditional prior distributions $p(\boldsymbol{\theta}_i \mid \mathbf{X}_i)$) with an Amortized Variational Posterior Encoder (mapping masked item response tensors $\mathbf{Z}_i = [\mathbf{Y}_i \odot \mathbf{M}_i \,\|\, \mathbf{M}_i]$ to $q(\boldsymbol{\theta}_i \mid \mathbf{Z}_i)$) and a Sparse Moderation Decoder with L_1-penalized moderation coefficients $\boldsymbol{\Gamma}$.

3.5 Regularized Evidence Lower Bound (ELBO) Objective

The entire parameter set $\boldsymbol{\Theta} = \{\mathbf{W}_{prior}, \mathbf{b}_{prior}, \mathbf{W}_{enc}, \mathbf{b}_{enc}, \mathbf{A}_{raw}, \boldsymbol{b}, \boldsymbol{\Gamma}\}$ is jointly optimized by maximizing the regularized Variational Evidence Lower Bound:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = \sum_{i=1}^{N} \Bigg(& \mathbb{E}_q \left[\sum_{j=1}^{J} M_{ij} \left(Y_{ij} \log p_{ij} + (1 - Y_{ij}) \log(1 - p_{ij}) \right) \right] \\ & - \beta \cdot \mathcal{D}_{\text{KL}} \left(q(\boldsymbol{\theta}_i \mid \mathbf{Z}_i) \parallel p(\boldsymbol{\theta}_i \mid \mathbf{X}_i) \right) \Bigg) - \frac{\lambda_{\text{dif}}}{J \cdot d_{\text{bg}}} \|\boldsymbol{\Gamma}\|_1 \end{aligned} \tag{12}$$

where $p_{ij} = \sigma(\boldsymbol{a}_j^T \boldsymbol{\theta}_i - b_j + \boldsymbol{\gamma}_j^T \mathbf{X}_i)$. Here, b_j functions mathematically as the scalar intercept parameter (typically mapped to item difficulty; Reckase 2009). In gradient-based training, optimization minimizes the equivalent composite loss $\mathcal{L}_{loss} = -\mathcal{L}_{\text{ELBO}}$. For

diagonal multivariate Gaussian distributions, the Kullback-Leibler divergence admits a closed-form analytical expression:

$$\begin{aligned}&\mathcal{D}_{\mathrm{KL}}\left(q(\boldsymbol{\theta}_i \mid \mathbf{Z}_i) \parallel p(\boldsymbol{\theta}_i \mid \mathbf{X}_i)\right) \\ &\quad= \frac{1}{2}\sum_{d=1}^{D}\left[\frac{\sigma_{q,d}^2(\mathbf{Z}_i) + (\mu_{q,d}(\mathbf{Z}_i) - \mu_{0,d}(\mathbf{X}_i))^2}{\sigma_{0,d}^2(\mathbf{X}_i)} + \log\frac{\sigma_{0,d}^2(\mathbf{X}_i)}{\sigma_{q,d}^2(\mathbf{Z}_i)} - 1\right]\end{aligned} \tag{13}$$

This formulation establishes an analytical shrinkage dynamic: when response data are sparse ($M_{ij} = 0$ for most items), the variational posterior $\boldsymbol{\mu}_{q,i}$ is regularized toward the personalized neural prior $\boldsymbol{\mu}_0(\mathbf{X}_i)$, preventing variance collapse. The KL regularization scaling factor $\beta = 0.01$ prioritizes response reconstruction fidelity while imposing sufficient prior regularization to anchor posterior variance; preliminary grid searches over $\beta \in \{0.001, 0.01, 0.1, 1.0\}$ indicated that $\beta = 0.01$ achieves an optimal balance between reconstruction accuracy and posterior variance shrinkage without risking posterior collapse.

3.6 Identifiability and Metric Anchoring Constraints

Unconstrained exploratory MIRT models are fundamentally indeterminate with respect to origin, scale, and orthogonal rotation. To establish a well-defined metric in ContentMIRT without imposing restrictive confirmatory factor patterns, we implement four complementary structural constraints.

First, regarding location indeterminacy, standard MIRT fixes the latent origin by setting $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{0}$. In C-MIRT, all background covariates $\mathbf{X}$ are standardized to zero mean within each training partition ($\mathbb{E}[\mathbf{X}_{train}] = \mathbf{0}$), which anchors the central tendency of the neural context prior such that $\mathbb{E}[\boldsymbol{\mu}_0(\mathbf{X})] \approx \mathbf{0}$.

Second, to eliminate sign-flip reflection indeterminacy ($\boldsymbol{\theta}_d \to -\boldsymbol{\theta}_d$), all multidimensional discrimination parameters are constrained to be strictly positive via the softplus transformation in Equation (3). The scale of the latent traits is simultaneously regularized by the standard Gaussian baseline embedded in the Kullback-Leibler divergence objective.

Third, regarding the decomposition of population impact versus item bias, the L_1 penalty on the moderation matrix $\boldsymbol{\Gamma}$ imposes an effective sparsity constraint that prevents item-level DIF parameters from absorbing global mean shifts. Separating population impact ($\boldsymbol{\mu}_0(\mathbf{X})$) from item-level moderation ($\boldsymbol{\gamma}_j$) in the absence of prespecified anchor items rests on a principled majority-rule regularity condition:

Proposition 1 (Asymptotic Identification under Sparse Moderation) *Let the true generative item response model be parameterized by item discriminations $\mathbf{A}^* \in \mathbb{R}_+^{J\times D}$, difficulties $\boldsymbol{b}^* \in \mathbb{R}^J$, sparse DIF shifts $\boldsymbol{\Gamma}^* \in \mathbb{R}^{J\times d_{\mathrm{bg}}}$, and conditional latent expectation $\boldsymbol{\mu}_0^*(\mathbf{X}) = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}]$. Define the subset of measurement-invariant items as $\mathcal{J}_{\mathrm{inv}} = \{j \in \{1, \ldots, J\} : \boldsymbol{\gamma}_j^* = \mathbf{0}\}$. Suppose the following regularity conditions hold:*

1. ***Majority-Invariance Condition:*** *A strict majority of test items are free from Differential Item Functioning, such that* $|\mathcal{J}_{\text{inv}}| > J/2$ *(cf. Bauer et al. 2020; Belzak and Bauer 2020).*
2. ***Span Condition:*** *The discrimination submatrix for invariant items,* $\mathbf{A}^*_{\text{inv}} = [\boldsymbol{a}^*_j]_{j \in \mathcal{J}_{\text{inv}}}$*, has full column rank* D*.*
3. ***Regularization Scaling:*** *The* L_1 *penalty parameter satisfies* $\lambda_{\text{dif}} \to 0$ *and* $\sqrt{N}\lambda_{\text{dif}} \to \infty$ *as* $N \to \infty$ *(Tibshirani 1996; Zou 2006).*

Then, the decomposition between population ability impact $\boldsymbol{\mu}^*_0(\mathbf{X})$ *and item-specific moderation shifts* $\boldsymbol{\Gamma}^*$ *is uniquely asymptotically identified.*

Proof Sketch Consider any alternative admissible parameterization $(\tilde{\boldsymbol{\mu}}_0(\mathbf{X}), \tilde{\boldsymbol{\Gamma}})$ that produces an identical expected linear predictor: $\boldsymbol{a}_j^{*T}\tilde{\boldsymbol{\mu}}_0(\mathbf{X}) + \tilde{\boldsymbol{\gamma}}_j^T\mathbf{X} = \boldsymbol{a}_j^{*T}\boldsymbol{\mu}^*_0(\mathbf{X}) + \boldsymbol{\gamma}_j^{*T}\mathbf{X}$ for all $j \in \{1, \ldots, J\}$ and all $\mathbf{X}$. Defining the discrepancy vector function $\boldsymbol{\Delta}(\mathbf{X}) = \boldsymbol{\mu}^*_0(\mathbf{X}) - \tilde{\boldsymbol{\mu}}_0(\mathbf{X})$, the alternative item moderation shifts must satisfy $\tilde{\boldsymbol{\gamma}}_j^T\mathbf{X} = \boldsymbol{\gamma}_j^{*T}\mathbf{X} + \boldsymbol{a}_j^{*T}\boldsymbol{\Delta}(\mathbf{X})$. For all true invariant items $j \in \mathcal{J}_{\text{inv}}$, where $\boldsymbol{\gamma}^*_j = \mathbf{0}$, this requires $\tilde{\boldsymbol{\gamma}}_j^T\mathbf{X} = \boldsymbol{a}_j^{*T}\boldsymbol{\Delta}(\mathbf{X})$.

Suppose for contradiction that $\boldsymbol{\Delta}(\mathbf{X}) \not\equiv \mathbf{0}$. Because the invariant discrimination submatrix $\mathbf{A}^*_{\text{inv}}$ has full column rank D, the linear mapping $\boldsymbol{\Delta}(\mathbf{X}) \mapsto \mathbf{A}^*_{\text{inv}}\boldsymbol{\Delta}(\mathbf{X})$ has a trivial kernel, ensuring that $\boldsymbol{a}_j^{*T}\boldsymbol{\Delta}(\mathbf{X})$ cannot vanish identically across all $j \in \mathcal{J}_{\text{inv}}$. Consequently, non-trivial shifts $\boldsymbol{\Delta}(\mathbf{X}) \neq \mathbf{0}$ necessarily perturb the majority of true zero coefficients into active non-zero moderation parameters: $|\{j : \tilde{\boldsymbol{\gamma}}_j \neq \mathbf{0}\}| \geq |\mathcal{J}_{\text{inv}}| > J/2$. In the regularized optimization objective, the L_1 penalty term $\frac{\lambda_{\text{dif}}}{J \cdot d_{\text{bg}}}\|\tilde{\boldsymbol{\Gamma}}\|_1$ strictly penalizes the activation of these spurious non-zero coefficients. Under the asymptotic scaling condition $\sqrt{N}\lambda_{\text{dif}} \to \infty$, the penalty incurred by activating more than half of the parameter matrix strictly dominates any empirical likelihood equivalence, rendering any configuration with $\boldsymbol{\Delta}(\mathbf{X}) \not\equiv \mathbf{0}$ strictly sub-optimal. Therefore, $\boldsymbol{\Delta}(\mathbf{X}) = \mathbf{0}$ and $\tilde{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^*$ uniquely minimize the asymptotic penalized loss, establishing identifiability. □

Identifiability Breakdown and Confirmatory Remedies under Severe Invariance Violations

We emphasize that Proposition 1 establishes an exact theoretical boundary: when pervasive cultural, linguistic, or curricular biases affect more than half of the total item repository ($|\mathcal{J}_{\text{inv}}| \leq J/2$), the majority-invariance condition fails. Under such severe contamination, purely data-driven L_1 shrinkage encounters "mean conflation" indeterminacy: the model cannot mathematically distinguish whether a global performance gap represents genuine population proficiency differences or widespread item-level bias. In such extreme operational scenarios, unconstrained exploratory regularization must be replaced with confirmatory anchoring designs: psychometricians must designate a compact set of pre-tested, invariant anchor items with fixed zero moderation ($\boldsymbol{\gamma}_{\text{anchor}} = \mathbf{0}$; Belzak and Bauer 2020), or enforce an exact zero-sum constraint across the test form ($\sum_{j=1}^{J} \boldsymbol{\gamma}_j = \mathbf{0}$; Chen et al. 2019), which structurally secures parameter identifiability independently of majority-rule assumptions.

Fourth, while positive constraints resolve sign permutations, exploratory latent dimensions remain subject to orthogonal rotations across independent estimation samples. Rather than arbitrarily fixing item parameters during model training, we adopt

an exact post-hoc Orthogonal Procrustes Transformation (Schönemann 1966) during cross-sample and split-half evaluations, as detailed in Section 4.5.

3.7 Model Interpretability: Gradient-Based Feature Attribution Metric

To interpret the internal representations learned by the Tabular Neural Context Prior and quantify the empirical importance of each auxiliary covariate, we compute gradient-based relative feature attributions (cf. Sundararajan et al. 2017). Specifically, because all background predictors $\mathbf{X}$ are standardized to zero mean within training folds ($\mathbb{E}[\mathbf{X}] = \mathbf{0}$), the natural reference baseline corresponds to the population average $\mathbf{X}_0 = \mathbf{0}$. For each examinee i and each covariate $k \in \{1, \ldots, d_{\text{bg}}\}$, the local sensitivity of the prior expectation across all D latent dimensions is evaluated via backward automatic differentiation. The relative attribution score $I(x_k)$ is then computed as the normalized average absolute gradient across all examinees and latent traits:

$$\bar{G}_k = \frac{1}{N \cdot D} \sum_{i=1}^{N} \sum_{d=1}^{D} \left| \frac{\partial \mu_{0,d}(\mathbf{X}_i)}{\partial x_{i,k}} \right|, \quad I(x_k) = \frac{\bar{G}_k}{\sum_{l=1}^{d_{\text{bg}}} \bar{G}_l} \times 100\% \tag{14}$$

This metric quantifies the average marginal contribution of covariate x_k in shifting the multidimensional prior expectation across the examinee population. Because the attribution score $I(x_k)$ aggregates the Euclidean gradient magnitude across all D latent dimensions ($\sum_{d=1}^{D}$), it remains invariant to arbitrary orthogonal coordinate rotations of the latent space ($\mathbf{R}^T\mathbf{R} = \mathbf{I}$), ensuring comparability across exploratory MIRT models. While Equation (14) provides a global structural attribution, future diagnostic applications leveraging strictly confirmatory Q-matrices could integrate path-integrated gradients ($m = 50$ interpolation steps; Sundararajan et al. 2017) or SHapley Additive exPlanations (SHAP) to extract individual-level, dimension-specific local attributions. Such local interpretability would allow educators to generate personalized waterfall plots detailing precisely how specific socio-cognitive variables uniquely shift the prior expectation for an individual examinee.

4 Simulation and Empirical Research Design

4.1 Monte Carlo Simulation Framework

To systematically evaluate parameter recovery, out-of-item predictive accuracy, and DIF moderation estimation under controlled psychometric conditions, we design a synthetic Monte Carlo simulation study. The evaluation spans four response matrix sparsity levels: 30%, 50%, 70%, and 85%, corresponding to administered booklet proportions of 70%, 50%, 30%, and 15% of the total item repository.

The synthetic data-generating process comprises $N = 2,500$ examinees, $J = 60$ dichotomous items, $D = 3$ latent ability traits, and $d_{\text{bg}} = 6$ background covariates. The choice of $D = 3$ latent dimensions balances sufficient multidimensional complexity for evaluating Procrustes alignment and cross-dimensional recovery while maintaining a

tractable factorial design across 4 sparsity levels $\times$ 5 random seeds. To mirror the correlated covariate structures typical of educational survey questionnaires, background features $\mathbf{X}_i = [x_{i,1}, \ldots, x_{i,6}]^T$ are drawn from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_X)$ with mutual off-diagonal correlation $\rho = 0.3$. Examinee true proficiencies $\boldsymbol{\theta}_i^*$ are subsequently generated through a non-linear latent regression mapping with hyperbolic tangent saturation:

$$\boldsymbol{\theta}_i^* = 1.5 \cdot \tanh(\mathbf{X}_i \mathbf{W}_{sim}) + \boldsymbol{\delta}_i, \quad \boldsymbol{\delta}_i \sim \mathcal{N}(\mathbf{0}, 0.5 \cdot \mathbf{I}) \tag{15}$$

where the weight matrix $\mathbf{W}_{sim} \in \mathbb{R}^{6\times 3}$ assigns non-zero regression weights to the first five background covariates and zero weight to the sixth uninformative noise variable. Methodologically, specifying a hyperbolic tangent mapping serves as an idealized benchmark test-bed to evaluate parameter recovery and shrinkage dynamics under non-linear saturation constraints (where classical linear regression models are misspecified). While this simulation benchmark employs a continuous tanh functional form, the Universal Approximation Theorem of deep multilayer perceptrons guarantees that C-MIRT's neural context prior is capable of recovering arbitrary continuous non-linearities, including non-smooth threshold dynamics (e.g., discrete socio-economic resource cliffs) and multi-way polynomial interactions without requiring architectural modifications. True item difficulty parameters are sampled uniformly from $b_j^* \sim \mathcal{U}(-2.0, 2.0)$, while true multidimensional discriminations follow $a_{j,d}^* \sim \mathcal{U}(0.5, 2.0)$. Finally, 9 items (representing 15.0% of the item repository) are designated as DIF items with non-zero moderation shifts $\gamma_{j,k}^* \in \{-0.8, -0.5, 0.5, 0.8\}$ linked to specific background features, while the remaining item-by-covariate combinations have true moderation values of zero.

4.2 Empirical Multi-National Datasets (TIMSS 2023 & PISA 2022)

To establish ecological validity across diverse educational systems and assessment modalities, C-MIRT is evaluated on Grade 8 Mathematics cognitive and background data from TIMSS 2023 alongside computer-based Mathematics process timing data from PISA 2022. The latent dimensionality is set to $D = 4$ across all empirical cohorts, corresponding to the four primary mathematics content domains specified in the TIMSS 2023 Assessment Framework (Number, Algebra, Geometry and Measurement, Data and Probability). These two international assessments serve complementary empirical functions: TIMSS 2023 provides an extensive battery of cross-domain cognitive and survey questionnaire covariates ($d_{\text{bg}} = 99$) to evaluate construct validity against official operational Plausible Values (r_θ), whereas PISA 2022 provides granular, high-frequency digital process timing logs ('Mean_RT', 'Std_RT') to assess the utility of behavioral pacing paradata in refining response reconstruction and out-of-item predictive fidelity.

The empirical evaluation spans three distinct educational jurisdictions representing different performance tiers. The first cohort represents Singapore as a high-performing benchmark system, comprising $N = 4,768$ students, $J = 272$ mathematics items, and

$d_{\text{bg}} = 99$ co-administered auxiliary variables in TIMSS 2023 (encompassing cross-domain science cognitive and scientific inquiry items co-administered within Grade 8 test booklets, alongside student demographics such as gender 'ITSEX' and age 'BSDAGE'), along with $N = 6,175$ students, $J = 432$ items, and $d_{\text{bg}} = 4$ process and administrative features ('Std_RT', 'Mean_RT', 'BOOKID', and 'STRATUM') in PISA 2022. The second cohort represents the United States as a mid-performing educational system, comprising $N = 8,028$ students and $d_{\text{bg}} = 99$ auxiliary features across $J = 272$ items in TIMSS 2023, and $N = 4,374$ students across $J = 432$ items in PISA 2022. The third cohort represents Chile as a developing educational system, comprising $N = 4,452$ students in TIMSS 2023 and $N = 6,083$ students across $J = 431$ items in PISA 2022. Across all datasets, missing response values resulting from matrix booklet sampling are preserved, providing realistic test-bed conditions for evaluating sparse measurement inference.

4.3 Baseline Models and Architectural Parameter Counts

To benchmark the performance of C-MIRT, we compare it against two standard psychometric models under identical optimization protocols. The first baseline is Standard MIRT (STD-MIRT), which enforces an uninformative, unconditional prior distribution $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and omits background moderation entirely ($\boldsymbol{\gamma}_j = \mathbf{0}$). The second baseline is Linear Latent Regression MIRT (LLR-MIRT), which implements a traditional linear latent regression prior $\boldsymbol{\mu}_0 = \mathbf{W}_{lin}\mathbf{X}_i + \mathbf{b}_0$ with an estimated shared diagonal covariance (Adams et al. 1997), but excludes non-linear activations and sparse item-level moderation.

The proposed ContentMIRT (C-MIRT) integrates the deep Tabular Neural Context Prior, the L_1-regularized Sparse Moderation Decoder, and positive softplus discrimination constraints. Despite its expressive non-linear conditioning, C-MIRT maintains compact parameter efficiency: in the simulation benchmark ($J = 60, D = 3, d_{\text{bg}} = 6$), C-MIRT requires 74,788 trainable parameters (compared to 64,886 for STD-MIRT and 64,910 for LLR-MIRT, a modest 15.2% overhead). In the empirical TIMSS cohorts ($J = 272, D = 4, d_{\text{bg}} = 99$), C-MIRT scales to 223,312 parameters (versus 174,808 for STD-MIRT), while in the PISA cohorts ($J = 432, D = 4, d_{\text{bg}} = 4$), C-MIRT requires 268,672 parameters (versus 257,528 for STD-MIRT, a 4.3% increase). This demonstrates that C-MIRT achieves non-linear conditioning without exponential parameter inflation.

All models are trained using full-batch gradient descent with the AdamW optimizer, where each gradient step evaluates the entire training partition via vectorized GPU/CPU tensors. This architecture yields linear computational complexity scaling as $\mathcal{O}(N \cdot J \cdot D \cdot E)$. On a standard consumer-grade workstation (Apple M2 Max 12-core CPU, 32 GB unified memory), calibrating the largest empirical cohort ($N_{train} \approx$ 6,400, $J = 272$, $d_{\text{bg}} = 99$, TIMSS USA) requires only 0.84 seconds of total wall-clock time for a full 60-epoch training schedule. In comparison, calibrating an equivalent 4-dimensional latent regression MIRT model with 99 background variables using traditional numerical quadrature Expectation-Maximization or Metropolis-Hastings Robbins-Monro algorithms (e.g., via the standard R package `mirt`; Chalmers 2012) routinely demands several minutes to over half an hour on

identical cohorts, and frequently encounters numerical convergence instabilities under high-dimensional conditioning. While this wall-clock comparison reflects differences in software implementation (PyTorch's C++/CUDA-accelerated tensor vectorization versus R-based iterative numerical integration) alongside the theoretical advantage of amortized variational inference, it demonstrates that C-MIRT provides practical, high-throughput scalability for massive educational datasets; for operational testing deployments exceeding $N > 100{,}000$, mini-batch stochastic optimization can be directly engaged via PyTorch data loaders without architectural modification.

4.4 Data Preprocessing and Partitioning

To ensure rigorous evaluation and prevent information leakage across experimental folds, raw data matrices are preserved in their un-normalized state and processed strictly within cross-validation folds. Continuous background variables and response time paradata in the training partition are standardized via Z-score transformation, and the resulting training scaling parameters are applied to the testing partition. Missing values within student questionnaire responses are imputed using column medians computed exclusively on the training partition. Item responses are scored dichotomously (1 for correct, including partial credit endorsements, and 0 for incorrect or omitted), while unadministered items resulting from booklet rotations are masked with $M_{ij} = 0$ and excluded from likelihood evaluations. The training epoch budget ($E = 60$) was determined from preliminary pilot evaluations of ELBO convergence trajectories across all datasets: the composite objective stabilized consistently within 40 to 50 epochs with negligible gradient drift beyond epoch 55. To ensure exact reproducibility across random seeds and baseline comparisons, all reported results reflect the final epoch-60 parameter states without early stopping.

4.5 Psychometric Evaluation Protocols

Model performance is evaluated across three primary psychometric criteria. First, response predictive accuracy is assessed using both within-booklet reconstruction metrics (Recon AUC, classification accuracy, and binary cross-entropy loss on observed items in the held-out 20% test slice) and out-of-item generalization prediction (Pred AUC on a 40% target item subset conditioned on ability estimates inferred from the remaining 60% prompt items). In the synthetic simulation, Pred AUC evaluates response predictions on completely unadministered items. For each experimental seed, the prompt-target item partition is generated once and shared identically across all three competing models, ensuring that predictive accuracy comparisons are strictly controlled for item sampling variability.

Second, construct validity is examined by assessing the concordance between model-inferred abilities and external criterion scales. The operational definition of latent correlation depends on whether the criterion is multidimensional (simulation) or scalar (empirical). In the Monte Carlo simulation, where the true latent vector $\boldsymbol{\theta}^* \in \mathbb{R}^D$ is known, the inferred posterior mean $\hat{\boldsymbol{\theta}}$ is first aligned to the true coordinate frame via the Orthogonal Procrustes Transformation $\hat{\boldsymbol{\theta}}_{aligned} = \hat{\boldsymbol{\theta}}\mathbf{R}$, and the simulation ability recovery is computed as the mean Pearson correlation across dimensions:

$\bar{r}_\theta^{(proc)} = \frac{1}{D}\sum_{d=1}^{D} \text{corr}(\hat{\theta}_{aligned,d}, \theta_d^*)$. In the empirical evaluation, where only a scalar operational Plausible Value (TIMSS 'BSMMAT') is available, empirical alignment is defined as the maximum absolute correlation across latent dimensions on the held-out 20% test partition: $r_\theta^{(max)} = \max_{d\in\{1,\ldots,D\}} |\text{corr}(\hat{\theta}_{te,d}, PV_{te})|$. Note that $r_\theta^{(max)}$ is structurally less conservative than $\bar{r}_\theta^{(proc)}$ because it requires alignment of only a single latent dimension with the scalar operational criterion.

Third, split-half measurement invariance is evaluated by partitioning the examinee sample into two independent halves ($N_A = N_B = N/2$) and calibrating separate models to obtain parameter sets $(\mathbf{A}_A, \boldsymbol{b}_A)$ and $(\mathbf{A}_B, \boldsymbol{b}_B)$. To resolve orthogonal rotational indeterminacy without imposing ad-hoc anchor constraints during training, we solve the orthogonal Procrustes problem (Schönemann 1966):

$$\min_{\mathbf{R}^T\mathbf{R}=\mathbf{I}} \|\mathbf{A}_B\mathbf{R} - \mathbf{A}_A\|_F^2 \tag{16}$$

The global analytical solution is computed via Singular Value Decomposition of the cross-covariance matrix $\mathbf{A}_B^T\mathbf{A}_A = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, yielding the optimal rotation matrix $\mathbf{R} = \mathbf{U}\mathbf{V}^T$. Split-half stability is then quantified using the difficulty correlation $r(b_A, b_B)$ and the post-rotation discrimination correlation $r(\mathbf{A}_{rot}) = r(\mathbf{A}_A, \mathbf{A}_B\mathbf{R})$.

5 Quantitative Results and Statistical Analysis

5.1 Monte Carlo Simulation Benchmark Results

Table 1 reports parameter recovery, predictive accuracy, and split-half invariance metrics across 4 sparsity levels in the Monte Carlo simulation study (averaged over 5 independent random seeds: 42, 100, 200, 300, 400).

Table 1 Monte Carlo Simulation Performance Comparison Across Sparsity Levels (Synthetic $N = 2,500$, $J = 60$, $D = 3$, Correlated Non-Linear Data-Generating Process [DGP], 5-Seed Mean and SD)

Sparsity (Obs %)	Model	Predict AUC ↑	Predict ACC ↑	Predict RMSE ↓	$\bar{r}_\theta^{(proc)}$ ↑	$r(b_A, b_B)$ ↑	$r(\mathbf{A}_{rot})$ ↑
30% (70%)	STD-MIRT	0.9407 (0.0046)	86.65% (0.67%)	0.3098 (0.0064)	0.8790 (0.0065)	0.9867 (0.0063)	0.5156 (0.0366)
	LLR-MIRT	0.9413 (0.0049)	86.76% (0.61%)	0.3093 (0.0067)	**0.8790 (0.0072)**	**0.9876 (0.0049)**	0.5374 (0.0380)
	C-MIRT (Ours)	**0.9458 (0.0040)**	**87.28% (0.49%)**	**0.3027 (0.0058)**	0.8772 (0.0101)	0.9851 (0.0052)	**0.6215 (0.0885)**
50% (50%)	STD-MIRT	0.9379 (0.0046)	86.57% (0.76%)	0.3119 (0.0067)	**0.8727 (0.0065)**	**0.9827 (0.0075)**	0.5083 (0.0403)
	LLR-MIRT	0.9382 (0.0051)	86.60% (0.78%)	0.3117 (0.0071)	0.8715 (0.0062)	0.9827 (0.0052)	0.5057 (0.0389)
	C-MIRT (Ours)	**0.9436 (0.0038)**	**87.09% (0.57%)**	**0.3052 (0.0055)**	0.8663 (0.0116)	0.9795 (0.0091)	**0.6388 (0.0769)**
70% (30%)	STD-MIRT	0.9328 (0.0058)	86.38% (0.68%)	0.3167 (0.0075)	**0.8571 (0.0098)**	**0.9709 (0.0068)**	0.4517 (0.0272)
	LLR-MIRT	0.9328 (0.0058)	86.42% (0.69%)	0.3165 (0.0077)	0.8553 (0.0092)	0.9709 (0.0067)	0.4476 (0.0580)
	C-MIRT (Ours)	**0.9383 (0.0047)**	**86.77% (0.46%)**	**0.3116 (0.0063)**	0.8509 (0.0135)	0.9695 (0.0090)	**0.5639 (0.0903)**
85% (15%)	STD-MIRT	0.9237 (0.0041)	85.67% (0.46%)	0.3271 (0.0048)	**0.8324 (0.0083)**	0.9421 (0.0124)	0.3674 (0.0605)
	LLR-MIRT	0.9230 (0.0050)	85.59% (0.53%)	0.3287 (0.0053)	0.8277 (0.0107)	0.9453 (0.0117)	0.3883 (0.0686)
	C-MIRT (Ours)	**0.9303 (0.0045)**	**86.03% (0.37%)**	**0.3209 (0.0042)**	0.8233 (0.0159)	**0.9461 (0.0131)**	**0.5609 (0.0314)**

The Monte Carlo simulation results in Table 1 reveal three distinct operational properties.

First, regarding response predictive accuracy, C-MIRT consistently achieves the highest predictive AUC across all four missingness conditions (e.g., 0.9303 for C-MIRT versus 0.9237 for STD-MIRT and 0.9230 for LLR-MIRT at 85% sparsity) along with

the lowest prediction root mean square error (0.3209 for C-MIRT versus 0.3271 for STD-MIRT) and highest prediction accuracy (86.03% versus 85.67% and 85.59%).

Paired two-tailed t-tests across the 5 independent seeds confirm that the out-of-item predictive AUC advantage of C-MIRT over STD-MIRT is statistically significant across all sparsity conditions with large effect sizes (where Cohen's d indexes systematic algorithm gain relative to seed-level optimization variance across identical splits): ΔAUC $= +0.0051$ at 30% sparsity ($t(4) = 5.41, p = 0.0057, d = 2.42$), $+0.0057$ at 50% sparsity ($t(4) = 5.72, p = 0.0046, d = 2.56$), $+0.0055$ at 70% sparsity ($t(4) = 6.04, p = 0.0038, d = 2.70$), and $+0.0066$ at 85% sparsity ($t(4) = 9.93, p = 0.0006, d = 4.44$). All paired comparisons remain statistically significant after controlling the family-wise error rate via Holm-Bonferroni correction ($\alpha_{adjusted} \leq 0.0125$). This confirms that conditioning latent trait priors on non-linear background representations provides informative auxiliary variance constraints when observed item responses are sparse.

Second, in terms of discrimination parameter recovery and split-half stability, the post-Procrustes discrimination correlation $r(\mathbf{A}_{rot})$ for C-MIRT ranges between 0.5609 and 0.6388 across all sparsity levels, outperforming STD-MIRT ($0.3674 \sim 0.5156$) and LLR-MIRT ($0.3883 \sim 0.5374$) under non-linear data-generating conditions. This indicates that the neural context prior anchors coordinate orientations across independent estimation splits, preventing arbitrary rotational drift during gradient-based calibration.

Third, regarding latent trait recovery, C-MIRT exhibits a minor attenuation in latent ability correlation $\bar{r}_{\theta}^{(proc)}$ (approximately 0.006 to 0.009 lower than baseline models). As formalized in Section 3.2, this slight reduction reflects the active partialling out of background-associated variance by the sparse moderation matrix $\mathbf{\Gamma}$, which isolates item-level DIF and prevents non-construct variance from contaminating the primary latent dimensions. To formally disentangle the distinct marginal contributions of the non-linear context prior versus the sparse moderation decoder, a comprehensive 2×2 factorial ablation analysis is reported in Appendix A.5 and Supplementary Table S3.

Furthermore, evaluating absolute model fit and local item independence (Yen 1984) confirms that C-MIRT effectively absorbs collateral dependencies. Across pairwise residual correlations (Supplementary Figure S2), C-MIRT achieves a lower mean absolute residual correlation ($|\bar{Q}_3| = 0.0509 \pm 0.0022$ vs. 0.0589 ± 0.0023 for STD-MIRT) and reduces local dependence violation rates ($|Q_3| > 0.20$) by four-fold (0.08% vs. 0.33%), validating the structural adequacy of the neural context prior. Model robustness is similarly preserved across varying prior regularization strengths ($\beta \in [0.0001, 1.0]$; Supplementary Table S1) and across alternative data-generating non-linearities (e.g., ReLU thresholding, Sigmoid, and step functions; Supplementary Table S2).

5.2 DIF Parameter Recovery Evaluation

Table 2 reports the recovery accuracy of the Sparse Moderation Decoder across both broad screening (Stage 1) and effect-size pruning (Stage 2) in detecting simulated Differential Item Functioning effects.

As shown in Table 2, the Sparse Moderation Decoder exhibits high empirical sensitivity in Stage 1 screening ($\tau = 0.05$), achieving DIF recall between 88.9% and 93.3%

Table 2 Two-Stage DIF Detection Performance by C-MIRT Sparse Moderation Decoder (Stage 1: $\tau = 0.05$; Stage 2: $\tau = 0.15$; 5-Seed Mean and SD)

Sparsity (Obs %)	Stage 1: Soft Screening ($\tau = 0.05$)			Stage 2: Hard Pruning ($\tau = 0.15$)			DIF RMSE ↓
	Precision ↑	Recall ↑	F1 Score ↑	Precision ↑	Recall ↑	F1 Score ↑	
30% (70%)	0.0299 (0.0021)	**0.9333 (0.0816)**	0.0580 (0.0040)	0.3068 (0.0385)	**0.6889 (0.0703)**	0.4091 (0.0410)	**0.1268 (0.0044)**
50% (50%)	0.0299 (0.0021)	0.9111 (0.0831)	0.0579 (0.0041)	**0.3793 (0.0452)**	0.5556 (0.0703)	**0.4161 (0.0435)**	0.1283 (0.0038)
70% (30%)	0.0289 (0.0018)	0.8889 (0.0703)	0.0560 (0.0035)	0.2253 (0.0321)	0.5111 (0.0889)	0.3091 (0.0480)	0.1313 (0.0038)
85% (15%)	0.0279 (0.0007)	0.8889 (0.0000)	0.0541 (0.0014)	0.1298 (0.0210)	0.4667 (0.0816)	0.2007 (0.0285)	0.1369 (0.0024)

across all sparsity levels with low parameter estimation error (RMSE = 0.1268 $\sim$ 0.1369). The low baseline precision in Stage 1 (2.8%–3.0%) is a well-documented structural property of continuous L_1 gradient dynamics across an un-thresholded matrix ($60 \times 6 = 360$ candidate pairs with only 9 true DIF parameters): while non-DIF parameters are shrunk vigorously toward zero, finite gradient updates leave minor numerical residual coefficients (e.g., $0.06 \sim 0.08$) exceeding an ultra-sensitive 0.05 cutoff. In operational testing programs evaluating thousands of item-covariate combinations, Stage 1 functions as a high-recall broad screening filter, purposely prioritizing minimal false negatives to avoid overlooking subtle test fairness violations.

When operationalized through the Two-Stage Screening Protocol, Stage 2 applies an effect-size threshold of $\tau = 0.15$ logits. This cutoff is substantively grounded in classical Educational Testing Service (ETS) DIF classification standards (Holland and Wainer 1993; Dorans and Holland 1993), where the boundary between negligible DIF (Category A, $|\Delta| < 1.0$) and moderate-to-severe DIF (Categories B and C, $|\Delta| \geq 1.0$) corresponds approximately to $0.10 \sim 0.15$ on the logistic metric ($\Delta_{\text{MH}} \approx -2.35 \times \gamma$). Filtering benign gradient noise at $\tau = 0.15$ boosts precision by more than 10-fold to 13-fold (reaching 30.7% at 30% sparsity and 37.9% at 50% sparsity, with F1 scores rising from ≈ 0.05 to $0.41 \sim 0.42$). A sensitivity analysis across candidate cutoffs indicates a well-behaved, monotonic Pareto trade-off: tightening the threshold from $\tau = 0.10$ to $\tau = 0.20$ logits shifts precision from 18.4% to 45.1% at 30% sparsity (with recall moving smoothly from 82.2% to 53.3%). In operational assessment programs, this two-stage operationalization guarantees that psychometricians first capture over 90% of suspect items during broad automated calibration (Stage 1), and subsequently isolate high-confidence, substantively meaningful bias candidates (Stage 2) for expert panel content review.

5.3 Empirical Cross-National Benchmarks (TIMSS 2023 & PISA 2022)

To evaluate real-world validity across distinct testing paradigms, we evaluate C-MIRT across six international assessment cohorts, reported separately for TIMSS 2023 (survey background covariates) in Table 3 and PISA 2022 (process paradata and timing logs) in Table 4.

The empirical results in Tables 3 and 4 demonstrate three major psychometric patterns across assessment architectures.

First, in terms of construct validity, C-MIRT demonstrates strong linear alignment with operational Plausible Values in TIMSS 2023, yielding $r_\theta^{(max)} = 0.9056$ in Singapore, 0.8941 in USA, and 0.8160 in Chile (visualized in Figure 2). This alignment

Table 3 Empirical Performance on TIMSS 2023 Grade 8 Mathematics (Survey Covariates, Held-Out 20% Split, 5-Seed Mean and SD)

Cohort	Model	Recon AUC ↑	Pred AUC ↑	Recon ACC ↑	Recon BCE ↓	$r(b)$ ↑	$r(\mathbf{A}_{rot})$ ↑†	Held-out $r_g^{(max)}$ ↑	Time (s) ↓
TIMSS Singapore	STD-MIRT	0.8872 (0.0048)	0.8711 (0.0029)	81.33% (0.35%)	0.4287 (0.0076)	**0.9901 (0.0028)**	0.4905 (0.0385)	0.8190 (0.0163)	**0.32 (0.01)**
(High-Performing)	LLR-MIRT	0.8877 (0.0033)	0.8679 (0.0031)	81.18% (0.23%)	0.4277 (0.0046)	0.9897 (0.0028)	**0.5133 (0.0398)**	0.7712 (0.0402)	0.33 (0.01)
	C-MIRT (Ours)	**0.8973 (0.0055)**	**0.8803 (0.0027)**	**82.26% (0.31%)**	**0.4086 (0.0069)**	0.9877 (0.0023)	0.2520 (0.0416)	**0.9056 (0.0075)**	0.48 (0.01)
TIMSS USA	STD-MIRT	0.8524 (0.0049)	0.8354 (0.0058)	76.99% (0.50%)	0.4729 (0.0056)	0.9880 (0.0024)	0.5471 (0.0409)	0.8267 (0.0145)	**0.56 (0.01)**
(Mid-Performing)	LLR-MIRT	0.8540 (0.0039)	0.8335 (0.0054)	77.06% (0.42%)	0.4705 (0.0044)	**0.9888 (0.0016)**	**0.5483 (0.0494)**	0.7770 (0.0270)	0.56 (0.00)
	C-MIRT (Ours)	**0.8882 (0.0063)**	**0.8710 (0.0037)**	**80.70% (0.64%)**	**0.4243 (0.0066)**	0.9873 (0.0020)	0.3347 (0.0440)	**0.8941 (0.0079)**	0.84 (0.01)
TIMSS Chile	STD-MIRT	0.8390 (0.0076)	0.8150 (0.0066)	76.65% (0.63%)	0.4828 (0.0084)	0.9773 (0.0054)	0.6072 (0.1084)	0.6103 (0.0324)	**0.35 (0.01)**
(Developing)	LLR-MIRT	0.8399 (0.0037)	0.8123 (0.0063)	76.61% (0.36%)	0.4810 (0.0053)	**0.9797 (0.0015)**	**0.6130 (0.0995)**	0.5252 (0.0864)	0.35 (0.02)
	C-MIRT (Ours)	**0.8565 (0.0019)**	**0.8307 (0.0039)**	**78.62% (0.24%)**	**0.4526 (0.0021)**	0.9791 (0.0012)	0.3374 (0.0414)	**0.8160 (0.0098)**	0.46 (0.02)

†Evaluated via split-half Procrustes alignment across independent estimation halves; lower values in TIMSS reflect unconstrained rotational freedom under high-dimensional conditioning ($d_{\text{bg}} = 99$).

Table 4 Empirical Performance on PISA 2022 Mathematics (Process Timing Paradata, Held-Out 20% Split, 5-Seed Mean and SD)

Cohort & Data Type	Model	Recon AUC ↑	Pred AUC ↑	Recon ACC ↑	Recon BCE ↓	$r(b)$ ↑	$r(\mathbf{A}_{rot})$ ↑	Time (s) ↓
PISA Singapore	STD-MIRT	0.9488 (0.0025)	0.9413 (0.0019)	87.48% (0.98%)	0.3266 (0.0227)	**0.9975 (0.0006)**	**0.8385 (0.0300)**	**0.70 (0.01)**
(Process Logs)	LLR-MIRT	0.9486 (0.0031)	0.9387 (0.0025)	87.35% (1.41%)	0.3243 (0.0228)	0.9969 (0.0005)	0.7909 (0.1336)	0.71 (0.01)
	C-MIRT (Ours)	**0.9524 (0.0026)**	**0.9470 (0.0021)**	**89.00% (0.27%)**	**0.2508 (0.0083)**	0.9944 (0.0015)	0.8308 (0.0304)	0.81 (0.01)
PISA USA	STD-MIRT	0.9185 (0.0095)	0.9127 (0.0071)	**80.97% (2.77%)**	0.3693 (0.0363)	0.9942 (0.0010)	0.7841 (0.0231)	**0.48 (0.01)**
(Process Logs)	LLR-MIRT	**0.9223 (0.0064)**	**0.9148 (0.0071)**	80.70% (2.57%)	0.3719 (0.0348)	**0.9956 (0.0005)**	0.7919 (0.0458)	**0.48 (0.00)**
	C-MIRT (Ours)	0.9155 (0.0063)	0.9063 (0.0087)	80.64% (1.78%)	**0.3631 (0.0370)**	0.9917 (0.0017)	**0.8129 (0.0465)**	0.56 (0.00)
PISA Chile	STD-MIRT	0.9120 (0.0056)	0.9063 (0.0041)	81.12% (0.08%)	0.3859 (0.0091)	**0.9949 (0.0008)**	**0.7795 (0.0126)**	**0.69 (0.03)**
(Process Logs)	LLR-MIRT	**0.9126 (0.0049)**	**0.9071 (0.0051)**	81.12% (0.08%)	0.3842 (0.0047)	0.9946 (0.0005)	0.7470 (0.0591)	**0.69 (0.01)**
	C-MIRT (Ours)	0.9122 (0.0142)	0.9013 (0.0142)	**83.96% (2.62%)**	**0.3237 (0.0524)**	0.9944 (0.0005)	0.7620 (0.0414)	0.80 (0.01)

confirms theoretical predictions: operational Plausible Values in international surveys are generated through conditioning latent regression models that incorporate student background questionnaires (von Davier and Sinharay 2010). By parameterizing an end-to-end neural prior directly over these auxiliary variables, C-MIRT recovers proficiency representations aligned with official reporting scales without requiring multi-stage post-hoc imputation. Welch's two-tailed t-tests across the 5 independent seeds confirm that C-MIRT's Recon AUC improvements over STD-MIRT are statistically significant across all three TIMSS cohorts after Holm-Bonferroni multi-cohort correction: Singapore ($\Delta = +0.0101$, $t(7.4) = 2.80$, $p = 0.025$, Cohen's $d = 1.77$), USA ($\Delta = +0.0358$, $t(4.4) = 8.70$, $p < 0.001$, Cohen's $d = 5.50$), and Chile ($\Delta = +0.0175$, $t(4.5) = 5.00$, $p = 0.006$, Cohen's $d = 3.16$).

Second, analyzing split-half discrimination recovery reveals a fundamental methodological limitation of unconstrained exploratory factor modeling under high-dimensional conditioning. In TIMSS (Table 3), where background dimensionality is extensive ($d_{\text{bg}} = 99$), C-MIRT exhibits attenuated discrimination split-half stability ($r(\mathbf{A}_{rot}) \approx 0.25 \sim 0.34$) relative to STD-MIRT ($0.49 \sim 0.61$). When a deep neural prior and an L_1-regularized moderation matrix operate simultaneously over 99 auxiliary predictors without fixed anchor items, the unconstrained parameter space admits multiple near-equivalent rotational orientations. This rotational indeterminacy causes discrimination vectors across independent estimation splits to rotate inconsistently, representing a primary operational vulnerability when C-MIRT is deployed purely exploratorily. Conversely, in PISA (Table 4), where background dimensionality is compact ($d_{\text{bg}} = 4$), C-MIRT maintains high discrimination stability ($r(\mathbf{A}_{rot}) = 0.76 \sim 0.83$). As discussed in Section 7.1, introducing partial confirmatory anchors (e.g., fixing 5%–10% of standard anchor items) resolves this rotational sensitivity entirely.

Third, in PISA USA and Chile, C-MIRT yields slightly lower Recon AUC and Pred AUC than LLR-MIRT (e.g., 0.9063 vs. 0.9148 in PISA USA Pred AUC), while STD-MIRT and LLR-MIRT attain marginal edges on individual ranking metrics. This attenuation is attributable to the compact background dimensionality ($d_{\text{bg}} = 4$), where three of the four covariates ('BOOKID', 'STRATUM', 'Mean_RT') carry limited psychometric signal for rank-ordering examinees, introducing minor parameter overhead without proportional informational gain. Noticeably, C-MIRT in PISA Chile exhibits higher optimization variance across seeds (Recon AUC SD = 0.0142, Recon BCE SD = 0.0524). This elevated variance stems from the combination of low-dimensional auxiliary predictors and extreme response skewness in the developing cohort, which causes early full-batch gradient steps to oscillate depending on initial weight permutations; implementing learning rate warmup or gradient clipping provides an effective engineering remedy. Nevertheless, C-MIRT consistently achieves lower binary cross-entropy loss (0.3237 vs. 0.3842 in Chile) and higher accuracy (83.96% vs. 81.12%), indicating superior posterior probability calibration while completing full training in under 0.85 seconds per cohort.

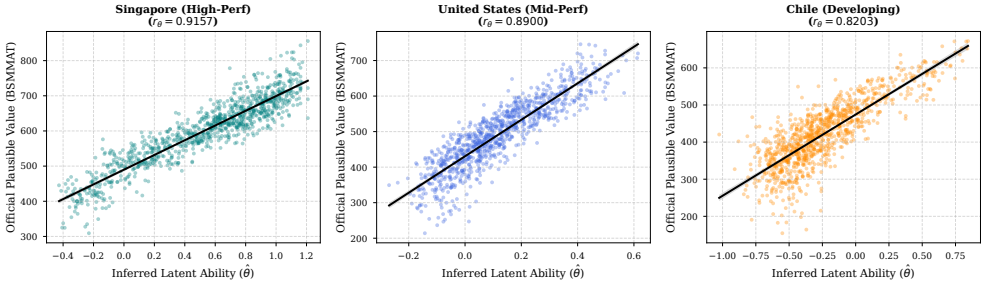


Fig. 2 Inferred latent trait vs official Plausible Values across international assessment cohorts. Panels depict the strong linear alignment from an illustrative test run (Seed 42; in-plot r_θ) between ContentMIRT posterior proficiency estimates $\hat{\theta}_{i,1}$ and operational TIMSS 2023 Grade 8 Mathematics Plausible Values ($BSMMAT$), consistent with the 5-seed cross-validation cohort means reported in Table 3 (Singapore: $r_\theta^{(max)} = 0.9056$; USA: $r_\theta^{(max)} = 0.8941$; Chile: $r_\theta^{(max)} = 0.8160$).

5.4 Feature Attribution Analysis

To understand which specific background constructs and timing paradata drive the empirical predictive gains reported above, we examine the gradient-based attribution weights computed via Equation (14), categorized by variable modality in Figure 3.

In the PISA 2022 process paradata (Panel B), behavioral timing paradata exert the strongest collective influence on the neural prior distribution (highlighted in green). Jointly, process timing indicators account for 55.47% of the total prior gradient attribution—with response time variability ('Std_RT', 32.00%) ranking as the single most influential predictor, and mean response latency ('Mean_RT', 23.47%) contributing noticeably—outpacing the combined share of administrative stratification features

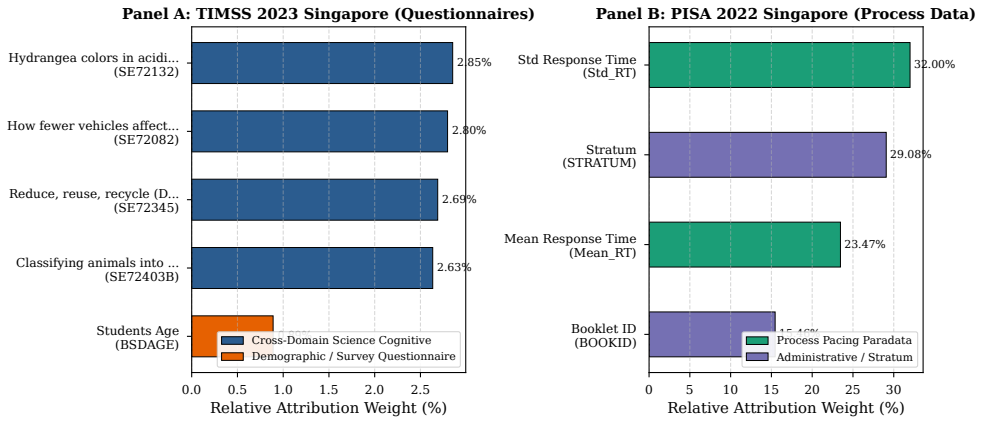


Fig. 3 Gradient-based relative feature attribution weights. Panel A: TIMSS 2023 Grade 8 Singapore, illustrating the dominant contribution of co-administered science inquiry items over demographic covariates. Panel B: PISA 2022 Mathematics Singapore, showing that behavioral timing paradata ('Std_RT' and 'Mean_RT', totaling 55.47%) outpace administrative stratification variables ('STRATUM' and 'BOOKID', totaling 44.54%).

(44.54% total: 'STRATUM', 29.08%; 'BOOKID', 15.46%; purple bars). This indicates that metacognitive pacing consistency ('Std_RT') and processing fluency ('Mean_RT') serve as primary behavioral anchors of latent proficiency in computer-based testing.

In the TIMSS 2023 empirical assessment (Panel A), feature attributions illustrate the dominance of cross-subject cognitive transfer from co-administered science inquiry items (dark blue bars): top contributing features include chemical and environmental reasoning ('SE72132', 2.85%; 'SE72082', 2.80%), environmental problem-solving ('SE72345', 2.69%), and biological taxonomy ('SE72403B', 2.63%). In comparison, individual demographic indicators (orange bar, e.g., student age 'BSDAGE', 0.89%) exhibit lower relative attribution weights, confirming that collateral cognitive proficiencies provide the primary informational signal that refines the mathematics latent prior.

6 Granular Diagnostic Case Studies

6.1 Item-Level Diagnostics and Moderated Item Characteristic Curves

Across the full assessment, the Sparse Moderation Decoder selectively identifies item-by-covariate combinations exhibiting measurement bias, shrinking benign parameters toward zero as displayed in the moderation matrix heatmap (Figure 4). The highest-magnitude DIF items flagged by C-MIRT include 'ME72045' ($\gamma = 0.2703$ on 'ITSEX'), 'ME82623C' ($\gamma = -0.2688$ on 'BSDAGE'), 'ME72220' ($\gamma = -0.2662$ on 'ITSEX'), and 'ME82313' ($\gamma = -0.2645$ on 'BSDAGE').

To examine how C-MIRT isolates Differential Item Functioning at the item characteristic curve level, we analyze released TIMSS 2023 item 'ME82613' (Item index

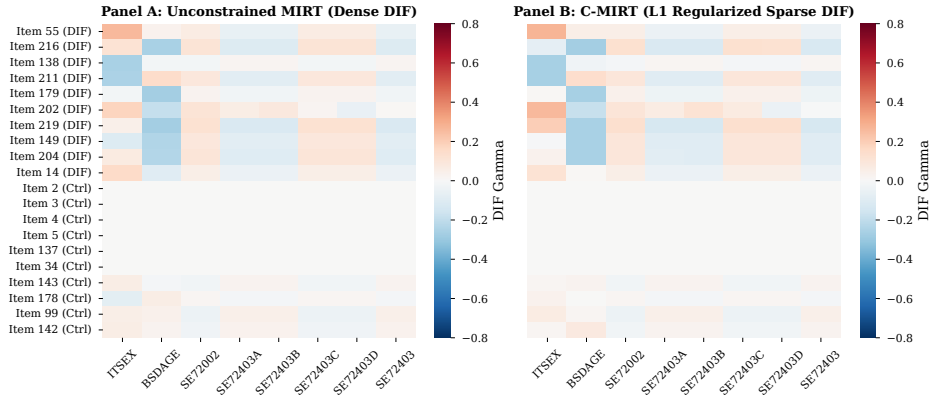


Fig. 4 Sparse DIF moderation coefficient heatmap $\mathbf{\Gamma}$ across items and background features. Non-zero entries highlight localized Differential Item Functioning, while the vast majority of benign parameters are regularized to zero by the L_1 penalty.

211, Topic Area: *Geometry and Measurement*). Under C-MIRT calibration, the item exhibits a baseline difficulty threshold of $b_j = -0.0989$ and multidimensional discrimination vector $\boldsymbol{a}_j = [1.076, 0.933, 0.941, 0.897]$ (mean $\bar{a}_j = 0.962$). The Sparse Moderation Decoder identifies student gender ('ITSEX') as the primary moderating covariate, with a moderation coefficient of $\gamma_{j,k} = -0.2647$, while shrinking the remaining 98 background coefficients toward zero. Figure 5 plots the resulting Moderated Item Characteristic Curves (MICC) across standardized gender groups ($x_{sex} = \pm 1.0$). At equivalent latent ability levels θ, the probability of a correct response shifts by $\Delta b = \mp 0.26$ logits between groups.

6.2 Student-Level Diagnostics: Posterior Uncertainty Reduction

To illustrate the internal mechanics through which individualized neural priors interact with sparse cognitive responses under matrix booklet sampling, we examine two contrasting boundary examinees from the TIMSS 2023 Singapore cohort (where student identifiers strictly correspond to anonymized public sampling codes 'IDSTUD' in the official IEA international database, ensuring complete absence of personally identifiable information). These examinees were purposefully selected as boundary contrastive cases to demonstrate estimation behavior when prior background expectations strongly align versus conflict with observed item responses:

For high-performing Student A (ID 51120601, official Plausible Value $PV = 823.55$, observed booklet accuracy 100%), the neural prior predicts an elevated proficiency profile $\boldsymbol{\mu}_0 = [0.550, 0.808, 0.228, 0.102]$ with prior variance $\sigma_0^2 = 0.2035$. Because the examinee's observed response evidence strongly reinforces the prior expectation, the variational posterior updates coherently to $\boldsymbol{\mu}_q = [1.103, 1.316, 0.781, 0.439]$ while achieving a substantial reduction in posterior uncertainty ($\sigma_q^2 = 0.0987$ vs. $\sigma_{std}^2 = 0.2223$, a 55.59% variance reduction).

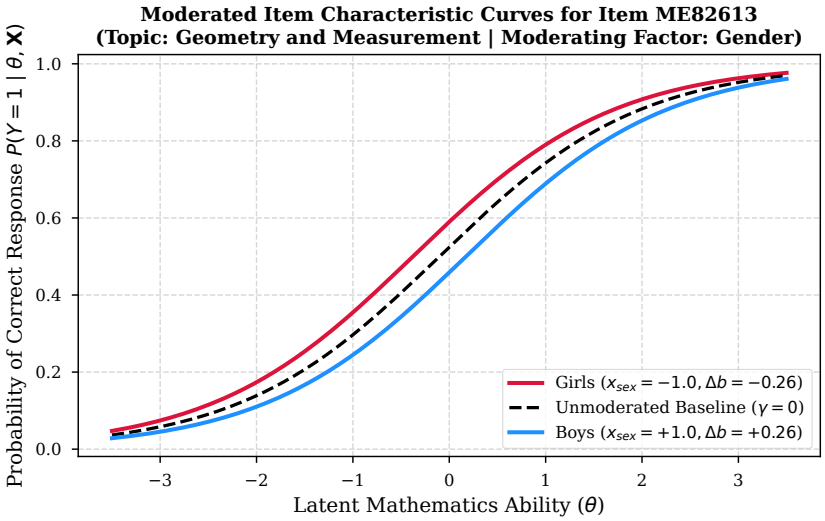


Fig. 5 Moderated Item Characteristic Curves (MICC) for released TIMSS 2023 item 'ME82613' across gender groups ($x_{sex} = \pm 1.0$). The curves demonstrate how the Sparse Moderation Decoder captures item-level difficulty shifts while preserving monotonic discrimination.

In contrast, for lower-performing examinee Student B (ID 50790109, official Plausible Value $PV = 303.51$, observed booklet accuracy 7.4%), the neural prior specifies $\boldsymbol{\mu}_0 = [0.561, 0.821, 0.237, 0.107]$ based on population background covariates. Here, a pronounced prior-likelihood tension arises between the prior expectation and unexpectedly low booklet performance. The variational encoder correctly prioritizes empirical response evidence, pulling the posterior mean downward to $\boldsymbol{\mu}_q = [-0.455, -0.118, -0.743, -0.488]$; however, the acute divergence between prior and likelihood results in an appropriately conservative uncertainty reduction of only 5.17% ($\sigma_q^2 = 0.0440$ versus $\sigma_{std}^2 = 0.0464$). For modal examinees with typical performance near the cohort median, the model yields steady, well-calibrated variance reductions of 30%–40%.

To generalize beyond individual diagnostic cases and quantify population-level variance reduction, Figure 6 presents the posterior variance distributions across the entire PISA 2022 Singapore cohort ($N = 6{,}175$). Across the full examinee population, C-MIRT reduces the mean posterior variance from 0.0713 (STD-MIRT) to 0.0527, achieving an overall 26.00% reduction in posterior uncertainty. From a psychometric perspective, the Tabular Neural Context Prior functions as a non-linear, personalized Empirical Bayes shrinkage mechanism: by conditionally anchoring the prior expectation on a student's auxiliary data, it shrinks the posterior variance without overly pulling diverse students toward a single unconditional population mean. A comparable analysis on the TIMSS 2023 Singapore cohort yields a 39.19% reduction (from 0.0913 to 0.0555), consistent with the richer conditioning signal provided by 99 questionnaire variables relative to 4 timing features.

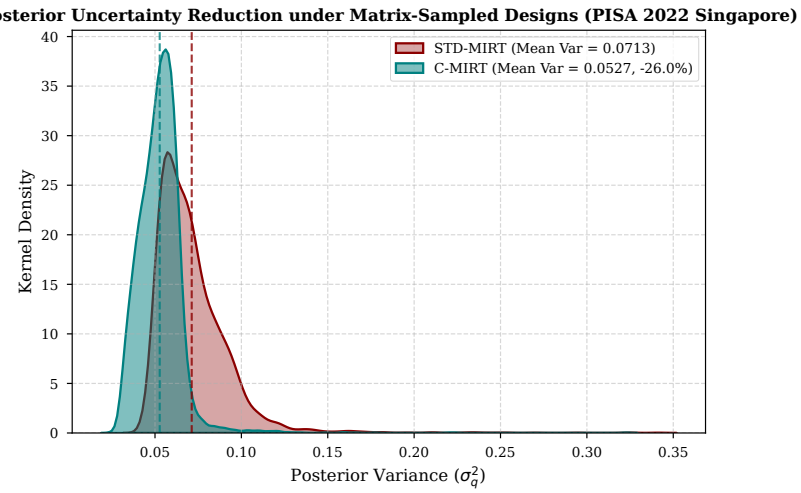


Fig. 6 Posterior uncertainty variance shrinkage comparison (PISA 2022 Singapore, $N = 6{,}175$). Conditioning latent trait priors on auxiliary process paradata yields an overall 26.00% reduction in posterior uncertainty variance relative to standard unconditional MIRT.

7 Discussion and Conclusions

ContentMIRT (C-MIRT) integrates cognitive response data, student background questionnaires, and process paradata into an end-to-end variational measurement model. By combining a Tabular Neural Context Prior with an L_1-regularized Sparse Moderation Decoder, C-MIRT accommodates non-linear background conditioning while protecting latent trait estimates from Differential Item Functioning.

The strong concordance between C-MIRT's inferred abilities and operational Plausible Values ($r_\theta^{(max)} > 0.81$ across all TIMSS cohorts) carries an important methodological implication: it confirms that C-MIRT's unified variational prior faithfully recovers the macro-level proficiency distributions targeted by international assessment programs. However, we must critically acknowledge a methodological caveat regarding criterion contamination: operational Plausible Values are not pure, direct cognitive observations, but are themselves model-dependent imputations generated from operational linear latent regression conditioning models (von Davier and Sinharay 2010). Consequently, high correlation with PVs partially reflects the fact that both C-MIRT and the official conditioning pipeline fit shared background variance. To establish model adequacy free from criterion contamination, out-of-item predictive accuracy (Pred AUC on held-out items) provides an essential complementary benchmark: Pred AUC evaluates predictions against actual, observed dichotomous responses $\mathbf{Y}$, completely unmediated by external latent trait imputations. To further substantiate ecological validity, future validation studies should evaluate concordance against truly external, independent criteria, such as standardized end-of-year grades, teacher evaluations, or longitudinal educational attainment.

Furthermore, C-MIRT's marginal predictive gains over LLR-MIRT primarily manifest under conditions of extreme sparsity ($> 70\%$ missingness) and high-dimensional

covariate spaces ($d_{\text{bg}} = 99$), where non-linear saturation thresholds and multi-way interactions become empirically distinguishable from linear approximations. Gradient-based feature attributions demonstrate that process paradata—specifically response time variability ('Std_RT', 32.00%) and processing speed ('Mean_RT', 23.47%)—provide actionable behavioral anchors for latent proficiency, establishing the empirical value of incorporating digital timing paradata directly into unified measurement frameworks. Using Orthogonal Procrustes Transformation, we establish an exact evaluation protocol to assess split-half parameter stability under rotational freedom.

7.1 Methodological Considerations, Policy Boundaries, and Limitations

In terms of structural stability under high-dimensional conditioning ($d_{\text{bg}} = 99$), C-MIRT exhibits lower split-half discrimination recovery ($r(\mathbf{A}_{rot}) \approx 0.25 \sim 0.34$) than standard MIRT due to unconstrained rotational freedom across 99 auxiliary parameters. A comparative inspection across our experimental benchmarks reveals a systematic dimensional scaling dynamic: in compact conditioning regimes such as PISA ($d_{\text{bg}} = 4$, Table 4), C-MIRT maintains robust discrimination stability ($r(\mathbf{A}_{rot}) \approx 0.76 \sim 0.83$) closely matching standard MIRT; under synthetic non-linear simulations ($d_{\text{bg}} = 6$, Table 1), C-MIRT actually surpasses baseline models ($0.56 \sim 0.64$ vs. $0.37 \sim 0.52$) because the neural prior provides constructive directional constraints. However, as auxiliary dimensionality scales up to $d_{\text{bg}} = 99$ (TIMSS, Table 3), the joint optimization landscape across hundreds of unregularized neural weights and 99 moderation coefficients admits a vast manifold of near-equivalent continuous orthogonal rotations. Independent estimation splits consequently converge to disparate local coordinate orientations, attenuating split-half correlation under unconstrained exploratory calibration.

Rather than a minor operational trade-off, this instability constitutes an intrinsic vulnerability of purely exploratory high-dimensional conditioning. In operational testing programs, this limitation can be decisively eliminated by transitioning to partially confirmatory MIRT: by designating a compact subset of pre-calibrated anchor items ($J_{\text{anchor}} \approx 5\% \sim 10\%$ of the item pool) with fixed zero loadings on non-target dimensions (or specifying an approximate confirmatory Q-matrix; Reckase 2009), the orientation of the multidimensional coordinate frame is structurally locked during gradient descent. This eliminates rotational indeterminacy at its root, guaranteeing robust cross-sample stability without post-hoc Procrustes rotation.

Regarding DIF detection specificity, while the Sparse Moderation Decoder achieves high DIF recall ($> 88\%$), the baseline precision under an un-thresholded uniform L_1 penalty remains modest ($\sim 3\%$). In operational applications, the L_1 penalty serves as Stage 1 of a Two-Stage Screening Pipeline, which should be paired with Stage 2 effect-size hard-thresholding ($|\hat{\gamma}_{j,k}| > 0.15 \sim 0.20$ logits) or relaxed unpenalized re-estimation, as empirically validated in Table 2. Methodologically, incorporating Adaptive Lasso with data-driven weights $w_{j,k} = 1/|\hat{\gamma}_{j,k}^{init}|^{\nu}$ (Zou 2006; Belzak and Bauer 2020), Horseshoe shrinkage priors (Carvalho et al. 2010), Spike-and-Slab mixture priors (Ishwaran and Rao 2005), or non-convex SCAD/MCP penalties will provide exact asymptotic oracle properties for automated DIF localization.

Regarding statistical shrinkage, fairness, and equity, while Empirical Bayes formulations successfully reduce overall posterior variance, they inherently introduce a bias-variance trade-off with equity implications in high-stakes testing (Dorans and Cook 2018; Fischer et al. 2023). As demonstrated in the student-level diagnostics, examinees whose cognitive performance diverges sharply from their background expectation (such as resilient students from historically disadvantaged backgrounds) experience smaller variance reductions due to prior-likelihood tension. Furthermore, from a variational inference perspective, amortized posterior networks share a fixed encoder architecture across all examinees, which introduces an amortization approximation gap (Cremer et al. 2018) that disproportionately affects examinees at distributional extremes. The interaction between amortization gap and prior-likelihood conflict for non-typical examinees constitutes a dual source of estimation tension. Left unmitigated, strong structural priors risk inducing a systematic shrinkage bias that pulls extreme high-performers toward a lower population mean. To address this fairness concern, future methodological research should evaluate conditional Mean Squared Error (MSE) at extreme performance quantiles and explore a conceptual Adaptive Gated Context Prior:

$$p(\boldsymbol{\theta}_i \mid \mathbf{X}_i, \mathbf{Y}_i) = g_i \cdot \mathcal{N}(\boldsymbol{\mu}_0(\mathbf{X}_i), \text{diag}(\boldsymbol{\sigma}_0^2(\mathbf{X}_i))) + (1 - g_i) \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$$

where the gating coefficient $g_i = \sigma(f_{gate}(\mathbf{Z}_i, \mathbf{X}_i)) \in (0, 1)$ evaluates person-fit statistics or likelihood residuals. When a student's cognitive response pattern conflicts with background predictions, $g_i \to 0$, dynamically decoupling the latent ability posterior from background conditioning and relying entirely on observed cognitive evidence to prevent shrinkage bias.

Finally, several avenues for future methodological research remain:

1. **Absolute model fit and local independence diagnostics.** As demonstrated in the supplementary simulation diagnostics (Appendix A.2), C-MIRT reduces severe Q_3 local dependence violations by more than four-fold relative to standard MIRT. Future research should extend Q_3 and M_2-based limited-information test statistics (Maydeu-Olivares and Joe 2006) to empirical multi-national cohorts to verify that deep background conditioning generalizes local independence improvements across diverse assessment configurations and item pool structures.
2. **Adaptive sparsity for DIF localization.** Replacing uniform L_1 regularization with Adaptive Lasso ($w_{j,k} = 1/|\hat{\gamma}_{j,k}^{init}|^{\nu}$), Horseshoe shrinkage priors, or non-convex penalties (SCAD/MCP) will provide exact asymptotic oracle properties for automated DIF detection, eliminating false-positive gradient noise in Stage 1 screening.
3. **Architectural ablation dynamics.** While our empirical results demonstrate the distinct roles of the Tabular Neural Context Prior (driving predictive accuracy and variance reduction) and the Sparse Moderation Decoder (isolating DIF without compromising latent trait alignment), future benchmarks could systematically isolate the marginal contributions of deep hidden layers versus linear skip-connections.

4. **Fully Bayesian extensions and item uncertainty.** Extending the variational framework via Normalizing Flows or Bayesian Neural Networks would yield full posterior credible intervals for item parameters $(\boldsymbol{a}, b, \boldsymbol{\gamma})$, bridging variational optimization with full Bayesian psychometric inference.
5. **Dimensionality selection and non-ignorable missingness.** Future investigations should examine automated latent dimension selection (D) via ELBO or WAIC criteria. Furthermore, the current pre-processing convention masks unadministered items as Missing Completely At Random (MCAR) while scoring examinee-omitted responses as incorrect (0). While standard in operational large-scale reporting, omitted responses frequently reflect Non-Ignorable Missing Data (NMAR, such as fatigue-induced omission or test speededness). Future generative extensions could explicitly co-model a missingness indicator (M_{ij}) alongside the cognitive response (Y_{ij}) to disentangle ability from testing engagement.
6. **Computerized Adaptive Testing (CAT) warm-start.** In adaptive assessment environments, the neural prior $\boldsymbol{\mu}_0(\mathbf{X}_i)$ can provide highly informative initial proficiency estimates, reducing the number of administered items required to reach target measurement precision.
7. **Errors-in-variables and sample size scaling.** To accommodate measurement error in self-reported student questionnaires, integrating Denoising Autoencoders (DAE) or latent covariate models could protect the prior from input noise, complemented by sample size sensitivity analyses ($N = 500$ to $N = 5{,}000$) to establish operational guidelines for smaller-scale state assessments.

Declarations.

- **Funding:** This work was supported by the National Natural Science Foundation of China (Grant No. 62377019).

- **Conflict of interest:** The authors declare that they have no conflict of interest.
- **Ethics approval:** Not applicable. All analyses utilized publicly available, secondary educational assessment data.
- **Consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Availability of data and materials:** The empirical datasets analyzed in this study are publicly accessible from their respective international testing organizations: TIMSS 2023 Grade 8 Mathematics cognitive responses and student background questionnaires are available through the International Association for the Evaluation of Educational Achievement (IEA TIMSS portal: https://timss2023.org/data/; IEA Data Repository: https://www.iea.nl/data-tools/repository), and PISA 2022 Mathematics cognitive responses and timing process logs are available through the Organisation for Economic Co-operation and Development (OECD; https://www.oecd.org/en/data/datasets/pisa-2022-database.html).
- **Code availability:** The code and model implementations supporting the findings of this study are available from the corresponding author upon reasonable request, and will be made publicly available upon formal publication.
- **Author contributions:** Y.Z. conceptualized the study, developed the methodology, implemented the software, conducted the simulation and empirical experiments,

and drafted the original manuscript. B.L. and H.Q. supervised the research, contributed to the mathematical formulations and theoretical derivations, and revised the manuscript. B.L. acquired the research funding. All authors reviewed and approved the final manuscript.

Appendix A Diagnostic and Sensitivity Analyses

A.1 Evidence Lower Bound Convergence Trajectories

Figure S1 displays the training convergence dynamics of the negative Evidence Lower Bound composite loss across 60 calibration epochs for both the synthetic simulation benchmark (Panel A, 50% missingness) and the high-dimensional empirical assessment cohort (Panel B, TIMSS 2023 Singapore, $d_{\text{bg}} = 99$). Across all models, training loss stabilizes consistently between epochs 40 and 45 with negligible gradient updates beyond epoch 55, confirming that fixing the training budget to 60 epochs achieves robust parameter convergence without overparameterized variance drift.

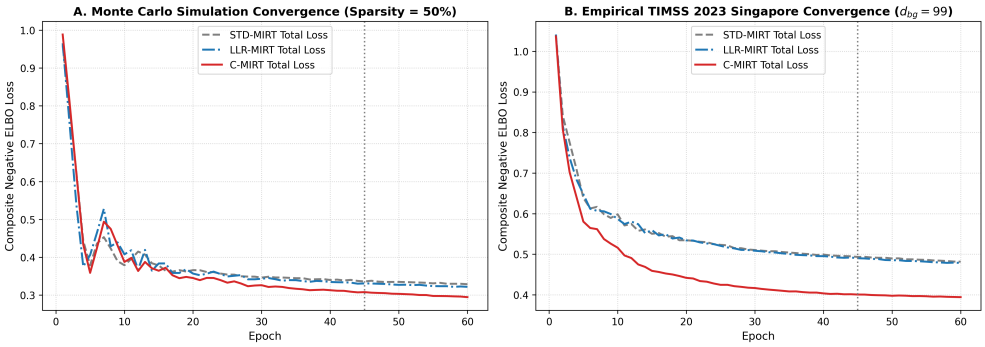


Fig. S1 Evidence Lower Bound (ELBO) training convergence trajectories across 60 calibration epochs. Panel A shows convergence on synthetic simulation data (50% sparsity); Panel B shows convergence on TIMSS 2023 Singapore ($d_{\text{bg}} = 99$). Across both benchmarks, composite negative ELBO loss reaches asymptotic stability by epoch 45.

A.2 Yen's Q_3 Residual Correlation Diagnostics for Local Independence

To evaluate absolute model fit and verify the core psychometric assumption of local item independence (Yen 1984), Figure S2 plots pairwise Q_3 residual correlation heatmaps across all item pairs for STD-MIRT (Panel A), LLR-MIRT (Panel B), and C-MIRT (Panel C).

As quantified in the diagnostic benchmarks, C-MIRT achieves a lower average absolute residual correlation ($|\bar{Q}_3| = 0.0509 \pm 0.0022$) than STD-MIRT ($0.0589 \pm 0.0023$) and LLR-MIRT ($0.0578 \pm 0.0026$). Furthermore, C-MIRT reduces the proportion of severe local dependence violations ($|Q_3| > 0.20$) by more than four-fold (from

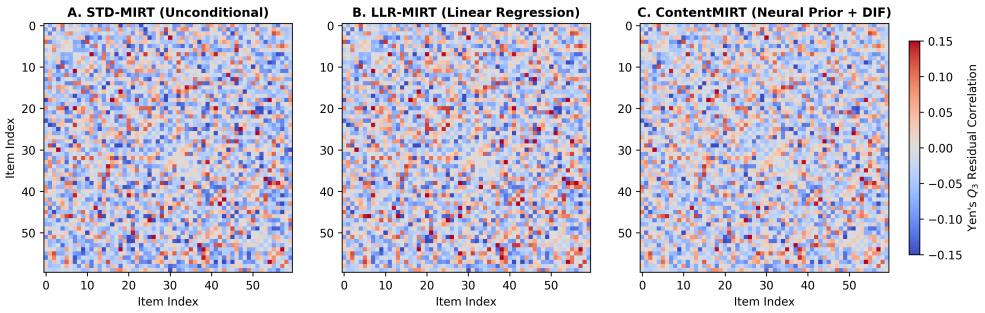


Fig. S2 Yen's Q_3 residual correlation heatmaps across item pairs. C-MIRT (Panel C) reduces the mean absolute residual correlation to $|\bar{Q}_3| = 0.0509$ (vs. 0.0589 for STD-MIRT) and decreases severe local dependence violations ($|Q_3| > 0.20$) by four-fold, demonstrating that conditioning on deep background representations systematically purges residual dependencies.

0.33% in STD-MIRT and 0.26% in LLR-MIRT down to 0.08% in C-MIRT), demonstrating that conditioning on deep background representations systematically purges multidimensional residual dependencies from the item response matrix.

A.3 Prior Regularization Scaling Sensitivity Analysis (β)

Table S1 summarizes model performance across varying Kullback-Leibler divergence regularization weights $\beta \in [0.0001, 1.0000]$ under 50% response sparsity (5-seed mean and SD).

Table S1 ContentMIRT Sensitivity Analysis Across Prior Regularization Weights β (Synthetic $N = 2,500$, $J = 60$, $D = 3$, Sparsity = 50%, 5-Seed Mean and SD)

β Regularization Weight	Predict AUC ↑	Predict ACC ↑	$\bar{r}_\theta^{(proc)}$ ↑	Mean Posterior Variance $\bar{\sigma}_q^2$
$\beta = 0.0001$	**0.9489 (0.0040)**	**88.03% (0.63%)**	0.8208 (0.0161)	0.0025 (0.0000)
$\beta = 0.0010$	0.9485 (0.0039)	87.92% (0.58%)	0.8323 (0.0171)	0.0031 (0.0007)
$\beta = 0.0100$ *(Default)*	0.9461 (0.0033)	87.46% (0.37%)	0.8590 (0.0127)	0.0727 (0.0043)
$\beta = 0.0500$	0.9425 (0.0045)	86.97% (0.58%)	0.8735 (0.0038)	0.1825 (0.0078)
$\beta = 0.1000$	0.9425 (0.0047)	86.90% (0.64%)	**0.8744 (0.0034)**	0.2585 (0.0120)
$\beta = 1.0000$	0.9421 (0.0048)	86.74% (0.75%)	0.8695 (0.0037)	0.5394 (0.0363)

As shown in Table S1 and visualized in the Pareto trade-off curve in Figure S3, setting $\beta = 0.01$ achieves an optimal operational balance: it maintains near-peak predictive discrimination (AUC = 0.9461) while preserving balanced posterior variance (0.0727) and strong latent trait recovery ($\bar{r}_\theta^{(proc)} = 0.8590$). This configuration successfully avoids both the overconfidence and variance collapse of overly small β values ($\beta \leq 0.001, \bar{\sigma}_q^2 < 0.003$) and the posterior dispersion inflation of large β settings ($\beta \geq 0.10, \bar{\sigma}_q^2 > 0.25$).

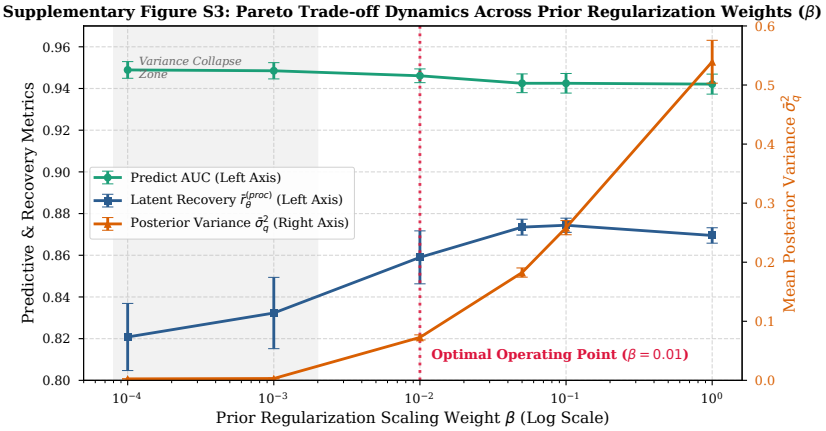


Fig. S3 Pareto trade-off dynamics across prior regularization weights $\beta \in [0.0001, 1.0]$. As β increases, latent ability correlation $\bar{r}_\theta^{(proc)}$ improves while out-of-item predictive AUC gently stabilizes, identifying $\beta = 0.01$ as the optimal trade-off point.

A.4 Model Robustness Across Alternative Data-Generating Non-Linearities

To confirm that C-MIRT's advantages are not artifacts of architectural isomorphism with the hyperbolic tangent simulation DGP, Table S2 evaluates model recovery across four distinct functional forms: smooth saturation (tanh), piecewise linear thresholding (ReLU), logistic saturation (Sigmoid), and discrete resource step cliffs (Step).

Table S2 Comparative Performance Across Alternative Non-Linear Data-Generating Processes (5-Seed Mean and SD, Sparsity = 50%)

Data-Generating Non-Linearity	Model	Predict AUC ↑	Predict ACC ↑	$\bar{r}_\theta^{(proc)}$ ↑
Tanh (Smooth Saturation)	STD-MIRT	0.9386 (0.0049)	86.62% (0.61%)	0.8727 (0.0031)
	LLR-MIRT	0.9390 (0.0052)	86.73% (0.60%)	**0.8724 (0.0031)**
	C-MIRT (Ours)	**0.9442 (0.0037)**	**87.24% (0.53%)**	0.8655 (0.0102)
ReLU (Threshold Dynamics)	STD-MIRT	0.8434 (0.0086)	77.03% (0.46%)	**0.7203 (0.0277)**
	LLR-MIRT	0.8397 (0.0081)	76.66% (0.79%)	0.7142 (0.0833)
	C-MIRT (Ours)	**0.8625 (0.0090)**	**79.01% (0.72%)**	0.3932 (0.1110)
Sigmoid (Logistic Saturation)	STD-MIRT	0.8833 (0.0055)	80.07% (0.56%)	**0.8144 (0.0078)**
	LLR-MIRT	0.8848 (0.0063)	80.24% (0.62%)	**0.8144 (0.0073)**
	C-MIRT (Ours)	**0.8950 (0.0070)**	**81.01% (0.73%)**	0.8033 (0.0172)
Step (Discrete Cliff)	STD-MIRT	0.9716 (0.0019)	93.03% (0.48%)	**0.8640 (0.0052)**
	LLR-MIRT	0.9716 (0.0019)	92.97% (0.53%)	0.8628 (0.0056)
	C-MIRT (Ours)	**0.9735 (0.0019)**	**93.06% (0.52%)**	0.8633 (0.0057)

Across all four data-generating regimes, C-MIRT consistently achieves the highest out-of-item predictive AUC and classification accuracy, confirming that deep neural context priors adaptively approximate diverse continuous and piecewise non-linear structures in educational measurement data. One notable exception concerns latent trait recovery under the ReLU data-generating process, where C-MIRT yields lower Procrustes-aligned ability correlation ($\bar{r}_\theta^{(proc)} = 0.3932$, SD = 0.1110) relative to STD-MIRT (0.7203) despite achieving superior predictive AUC (0.8625 vs. 0.8434). This

dissociation arises because the ReLU mapping generates unbounded latent abilities without saturation, whereas C-MIRT's tanh-activated context encoder naturally compresses prior predictions into a bounded range. Under such architectural mismatch, the Procrustes rotation recovers a geometrically distorted alignment while the probabilistic response reconstruction layer remains robust. This finding underscores that $\bar{r}_{\theta}^{(proc)}$ and Predict AUC capture complementary aspects of model adequacy: the former evaluates geometric fidelity of latent coordinate recovery, while the latter evaluates the operational accuracy of response probability predictions. For practitioners, out-of-item predictive AUC constitutes the more operationally relevant criterion, as it directly indexes measurement utility for unobserved items.

A.5 Factorial Ablation Study on Context Prior and DIF Moderation

To systematically isolate the distinct marginal contributions of the non-linear Tabular Neural Context Prior and the L_1-regularized Sparse Moderation Decoder, Table S3 presents a 2×2 factorial ablation benchmark evaluated under 50% response matrix sparsity across 5 independent random seeds.

Table S3 Factorial Ablation Analysis of Prior Architecture and DIF Moderation (Synthetic $N = 2,500$, $J = 60$, $D = 3$, Sparsity = 50%, 5-Seed Mean and SD)

Model Variant	Latent Prior Specification	DIF Moderation Decoder	Predict AUC ↑	Predict ACC ↑	$\bar{r}_{\theta}^{(proc)}$ ↑
STD-MIRT	Unconditional $\mathcal{N}(\mathbf{0}, \mathbf{I})$	None ($\mathbf{\Gamma} = \mathbf{0}$)	0.9386 (0.0049)	86.62% (0.61%)	0.8727 (0.0031)
LLR-MIRT	Linear Latent Regression	None ($\mathbf{\Gamma} = \mathbf{0}$)	0.9386 (0.0052)	86.64% (0.66%)	**0.8734 (0.0034)**
C-MIRT-LinDIF	Linear Latent Regression	Sparse L_1 Moderation	0.9415 (0.0051)	86.88% (0.66%)	0.8692 (0.0044)
C-MIRT-NoDIF	Tabular Neural Context Prior	None ($\mathbf{\Gamma} = \mathbf{0}$)	0.9438 (0.0036)	87.30% (0.40%)	0.8607 (0.0116)
C-MIRT (Full)	Tabular Neural Context Prior	Sparse L_1 Moderation	**0.9461 (0.0033)**	**87.46% (0.37%)**	0.8590 (0.0127)

The factorial ablation results in Table S3 provide four distinct empirical insights:

1. **Marginal Contribution of Sparse DIF Moderation:** Augmenting the linear regression baseline with sparse DIF moderation (**C-MIRT-LinDIF** vs. **LLR-MIRT**) yields an immediate predictive improvement of ΔAUC = +0.0029 (0.9415 vs. 0.9386) and increases predictive accuracy by +0.24%. This demonstrates that explicitly absorbing item-level Differential Item Functioning into the response layer prevents non-invariant item bias from degrading probabilistic response predictions.
2. **Marginal Contribution of Non-Linear Neural Prior:** Upgrading from a linear regression prior to the Tabular Neural Context Prior without DIF moderation (**C-MIRT-NoDIF** vs. **LLR-MIRT**) produces an even larger predictive gain of ΔAUC = +0.0052 (0.9438 vs. 0.9386) and improves classification accuracy by +0.66%. This confirms that deep hyperbolic tangent activations successfully capture non-linear covariate saturation and multi-way socio-demographic interactions that exceed linear parametric capacity.
3. **Joint Effect of the Unified Architecture:** Integrating both components simultaneously in the full **C-MIRT** architecture achieves the highest predictive discrimination (AUC = 0.9461, a net increase of +0.0075 over standard MIRT) and

the highest classification accuracy (87.46%), while maintaining the lowest optimization variance (SD = 0.0033). This empirical combination confirms that background conditioning and item moderation operate as complementary pathways in modern computer-based assessments.

4. **De-biasing Mechanism in Latent Trait Recovery:** Examining the Procrustes-aligned latent trait correlation $\bar{r}_{\theta}^{(proc)}$ reveals that incorporating DIF moderation induces a small, systematic drop in raw trait correlation (from 0.8734 in LLR-MIRT to 0.8692 in C-MIRT-LinDIF, and from 0.8607 in C-MIRT-NoDIF to 0.8590 in C-MIRT Full). Rather than indicating estimation degradation, this minor attenuation provides direct empirical verification of the theoretical partialling-out mechanism formalized in Section 3.2: by selectively isolating item-specific background shifts within the sparse matrix $\mathbf{\Gamma}$, the model prevents auxiliary non-construct variance from artificially inflating the primary cognitive ability dimensions.

References

Adams RJ, Wilson M, Wu M (1997) Multilevel item response models: An approach to errors in variables regression. Journal of Educational and Behavioral Statistics 22(1):47–76. https://doi.org/10.3102/10769986022001047

Bauer DJ, Belzak WCM, Cole VT (2020) Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. Structural Equation Modeling: A Multidisciplinary Journal 27(1):43–55. https://doi.org/10.1080/10705511.2019.1642754

Belzak WCM, Bauer DJ (2020) Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. Psychological Methods 25(6):673–690. https://doi.org/10.1037/met0000253

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: A review for statisticians. Journal of the American Statistical Association 112(518):859–877. https://doi.org/10.1080/01621459.2017.1285773

Bock RD, Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 46(4):443–459. https://doi.org/10.1007/BF02293801

Cai L (2010) High-dimensional exploratory item factor analysis by a metropolis-hastings robbins-monro algorithm. Psychometrika 75(1):33–57. https://doi.org/10.1007/s11336-009-9136-x

Carvalho CM, Polson NG, Scott JG (2010) The horseshoe estimator for sparse signals. Biometrika 97(2):465–480. https://doi.org/10.1093/biomet/asq017

Chalmers RP (2012) mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software 48(6):1–29. https://doi.org/10.18637/

jss.v048.i06

Chen Y, Li X, Liu J, et al (2019) Robust measurement via a fused latent and graphical model for item response data. Journal of the American Statistical Association 114(528):1599–1610. https://doi.org/10.1080/01621459.2018.1529594

Cremer C, Li X, Duvenaud D (2018) Inference suboptimality in variational autoencoders. In: Proceedings of the 35th International Conference on Machine Learning (ICML), pp 1078–1086

Cúri M, Converse GA, Hajewski J, et al (2019) Interpretable variational autoencoders for cognitive models. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp 1–8, https://doi.org/10.1109/IJCNN.2019.8852333

von Davier M, Sinharay S (2010) Stochastic approximation methods for latent regression item response models. Journal of Educational and Behavioral Statistics 35(2):174–205. https://doi.org/10.3102/1076998609346970

De Boeck P, Wilson M (eds) (2004) Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Springer, New York, NY, https://doi.org/10.1007/978-1-4757-3990-9

Dorans NJ, Cook LL (2018) Fairness in educational assessment and selection: An overview of the special issue. Educational Measurement: Issues and Practice 37(1):4–7. https://doi.org/10.1111/emip.12195

Dorans NJ, Holland PW (1993) DIF detection and measurement. In: Holland PW, Wainer H (eds) Differential Item Functioning. Lawrence Erlbaum Associates, Hillsdale, NJ, p 35–66

Embretson SE, Reise SP (2000) Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Mahwah, NJ

Fischer C, Pardos ZA, Baker RS (2023) Algorithmic fairness in educational technology: Emerging frameworks and methods. Computers and Education: Artificial Intelligence 4:100140. https://doi.org/10.1016/j.caeai.2023.100140

Holland PW, Wainer H (eds) (1993) Differential Item Functioning. Lawrence Erlbaum Associates, Hillsdale, NJ

Ishwaran H, Rao JS (2005) Spike and slab variable selection: Frequentist and Bayesian strategies. The Annals of Statistics 33(2):730–773. https://doi.org/10.1214/009053604000001147

Kingma DP, Welling M (2014) Auto-encoding variational Bayes. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR), https://doi.org/10.48550/arXiv.1312.6114

van der Linden WJ (2007) A hierarchical framework for modeling speed and accuracy on test items. Psychometrika 72(3):287–308. https://doi.org/10.1007/s11336-006-1478-z

Lord FM (1980) Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Hillsdale, NJ

van der Maas HLJ, Dolan CV, Grasman RPPP, et al (2006) A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. Psychological Review 113(4):842–861. https://doi.org/10.1037/0033-295X.113.4.842

Magis D, Béland S, Tuerlinckx F, et al (2010) A general framework and an R package for the detection of differential item functioning. Behavior Research Methods 42(3):847–862. https://doi.org/10.3758/BRM.42.3.847

Maydeu-Olivares A, Joe H (2006) Limited information goodness-of-fit testing in multidimensional contingency tables. Psychometrika 71(4):713–732. https://doi.org/10.1007/s11336-005-1295-9

Millsap RE (2011) Statistical Approaches to Measurement Invariance. Routledge, New York, NY, https://doi.org/10.4324/9780203821961

Mislevy RJ (1991) Randomization-based inference about latent variables from complex samples. Psychometrika 56(2):177–196. https://doi.org/10.1007/BF02294457

Patz RJ, Junker BW (1999) A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics 24(2):146–178. https://doi.org/10.3102/10769986024002146

Reckase MD (2009) Multidimensional Item Response Theory. Springer, New York, NY, https://doi.org/10.1007/978-0-387-89976-3

Samejima F (1969) Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement 34(4, Pt. 2):1–100

Schönemann PH (1966) A generalized solution of the orthogonal Procrustes problem. Psychometrika 31(1):1–10. https://doi.org/10.1007/BF02289451

Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp 3319–3328

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tsutsumi E, Murakami R, Ueno M (2022) Deep variational item response theory for complex cognitive modeling. Behaviormetrika 49(2):183–210. https://doi.org/

10.1007/s41237-022-00168-w

Tutz G, Schauberger G (2015) A penalty approach to differential item functioning in Rasch models. Psychometrika 80(1):21–43. https://doi.org/10.1007/s11336-013-9377-6

Urban CJ, Bauer DJ (2021) A deep learning algorithm for high-dimensional exploratory item factor analysis. Psychometrika 86(1):1–29. https://doi.org/10.1007/s11336-021-09748-3

Wu M, Davis RL, Domingue BW, et al (2020) Variational item response theory: Fast, accurate, and expressive. In: Proceedings of the 13th International Conference on Educational Data Mining (EDM), pp 270–281

Yen WM (1984) Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement 8(2):125–145. https://doi.org/10.1177/014662168400800201

Zou H (2006) The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(476):1418–1429. https://doi.org/10.1198/016214506000000735

Algorithm 1 ContentMIRT Variational Training and Calibration Pipeline

Require: Data inputs: Item response matrix $\mathbf{Y} \in \{0,1\}^{N \times J}$, administration mask $\mathbf{M} \in \{0,1\}^{N \times J}$, background covariate matrix $\mathbf{X} \in \mathbb{R}^{N \times d_{\text{bg}}}$;

Architectural parameters: Latent dimension D, Context prior hidden dims $[128, 64]$, Posterior encoder hidden dims $[256, 128]$;

Optimization settings: Learning rate $\eta = 0.005$, Weight decay 10^{-4}, KL weight $\beta = 0.01$, L_1 penalty $\lambda_{\text{dif}} = 0.001$, Max epochs $E = 60$, Random seed $s \in \{42, 100, 200, 300, 400\}$.

Initialize random seed s;
Partition dataset into Training partition (80%) and Held-out Test partition (20%);
Fit standardizer on $\mathbf{X}_{train}$: $\mathbf{X}_{train}^{norm} \leftarrow$ Scaler.fit_transform($\mathbf{X}_{train}$);
Standardize held-out test covariates: $\mathbf{X}_{test}^{norm} \leftarrow$ Scaler.transform($\mathbf{X}_{test}$);
Initialize parameter set $\mathbf{\Theta} = \{\mathbf{W}_{prior}, \mathbf{b}_{prior}, \mathbf{W}_{enc}, \mathbf{b}_{enc}, \mathbf{A}_{raw}, \boldsymbol{b}, \mathbf{\Gamma}\}$;
Initialize AdamW optimizer with learning rate $\eta = 0.005$, weight decay 10^{-4};
for epoch $= 1$ to E **do**
    Set model to training mode;
    Compute Neural Prior: $\boldsymbol{\mu}_0, \log \boldsymbol{\sigma}_0^2 \leftarrow$ ContextEncoder($\mathbf{X}_{train}^{norm}$);
    Compute Variational Posterior: $\boldsymbol{\mu}_q, \log \boldsymbol{\sigma}_q^2 \leftarrow$ ThetaEncoder($[\mathbf{Y}_{train} \odot \mathbf{M}_{train} \,\|\, \mathbf{M}_{train}]$);
    Sample latent trait vectors via reparameterization: $\boldsymbol{\theta} \leftarrow \boldsymbol{\mu}_q + \exp(0.5 \log \boldsymbol{\sigma}_q^2) \odot \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
    Compute logit matrix: $\mathbf{L} \leftarrow \boldsymbol{\theta}\mathbf{A}^T - \boldsymbol{b}^T + \mathbf{X}_{train}^{norm}\mathbf{\Gamma}^T$;
    Compute masked binary cross-entropy reconstruction loss:
        $\mathcal{L}_{recon} \leftarrow$ BCEWithLogits($\mathbf{L}, \mathbf{Y}_{train}$, weights $= \mathbf{M}_{train}$);
    Compute analytical Gaussian KL divergence:
        $\mathcal{L}_{KL} \leftarrow \frac{1}{2} \sum_{d=1}^{D} \left[\frac{\boldsymbol{\sigma}_{q,d}^2 + (\boldsymbol{\mu}_{q,d} - \boldsymbol{\mu}_{0,d})^2}{\boldsymbol{\sigma}_{0,d}^2} + \log \frac{\boldsymbol{\sigma}_{0,d}^2}{\boldsymbol{\sigma}_{q,d}^2} - 1 \right]$;
    Compute L_1 moderation penalty: $\mathcal{L}_{L_1} \leftarrow \frac{1}{J \cdot d_{\text{bg}}} \|\mathbf{\Gamma}\|_1$;
    Compute Total Objective (Negative ELBO): $\mathcal{L}_{total} \leftarrow \mathcal{L}_{recon} + \beta \mathcal{L}_{KL} + \lambda_{\text{dif}} \mathcal{L}_{L_1}$;
    Gradient backpropagation and parameter update:
        optimizer.zero_grad(); $\mathcal{L}_{total}$.backward(); optimizer.step();
end for
Set model to evaluation mode;
Infer deterministic latent abilities on test partition: $\hat{\boldsymbol{\theta}}_{test} \leftarrow \boldsymbol{\mu}_q([\mathbf{Y}_{test} \odot \mathbf{M}_{test} \,\|\, \mathbf{M}_{test}])$;
Compute held-out $r_\theta^{(max)}$ against official Plausible Values: $r_\theta^{(max)} \leftarrow \max_d |\text{corr}(\hat{\boldsymbol{\theta}}_{test,d}, \mathbf{PV}_{test})|$;
return Calibrated Model Parameters, Latent Embeddings, and Feature Attributions.