DUALKT: DUAL-TIMESCALE KNOWLEDGE TRACING VIA ORTHOGONAL DECOMPOSITION AND TARGET-CONDITIONED FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge tracing (KT) predicts future performance from students' interaction histories. Accurate prediction requires modeling both long-term evolution across the full history and local fluctuations driven by recent interactions. Yet existing methods often encode both scales jointly or directly combine multiple encoders without explicitly modeling the geometric relationship between their representations. We propose DualKT, a dual-timescale model with orthogonal decomposition and target-conditioned fusion. DualKT constructs a structured question representation, encodes the full history into a long-term state with Mamba-2, and uses a target-querying hierarchical Transformer to extract a target-relevant short-term state from recent interactions and chunked full-prefix summaries. It decomposes the short-term state into a long-term-aligned component and an orthogonal residual, then fuses them with the long-term state conditioned on the target question. In five-fold question-level experiments on five public datasets, DualKT ranks among the top two methods in AUC on every dataset and first on four under fixed configurations. Ablations identify the structured question representation as the main source of improvement, with further gains from the dual-timescale branches and learned fusion. These results support explicit modeling and integration of long- and short-term cognitive information for target-specific prediction.

1 INTRODUCTION

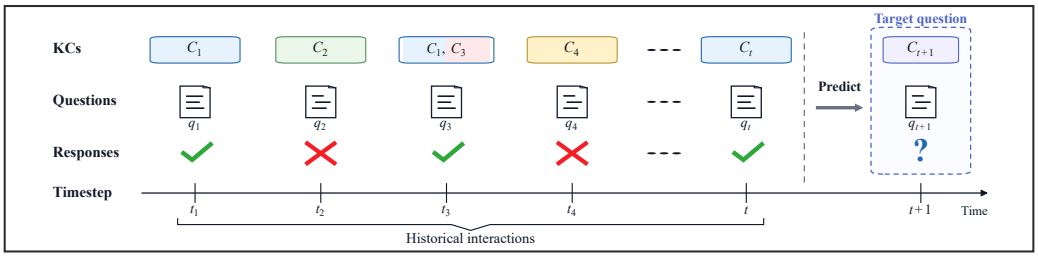
Online learning platforms and intelligent tutoring systems record large volumes of temporally ordered question–response interactions. Knowledge tracing (KT) uses these histories to estimate students' latent knowledge states and predict their responses to subsequent questions (Corbett & Anderson, 1995; Piech et al., 2015). Accurate KT supports personalized practice, learning diagnosis, and instructional intervention, making it a key component of adaptive educational systems.

To characterize how knowledge states evolve during learning, deep KT models have employed recurrent networks, memory networks, attention mechanisms, and state-space models. Recurrent and memory-based models preserve historical information by continually updating internal states; attention-based models retrieve predictive evidence according to the relevance between the target question and past interactions; and state-space models provide an efficient alternative for long-sequence modeling (Piech et al., 2015; Zhang et al., 2017; Pandey & Karypis, 2019; Ghosh et al., 2020; Cao & Zhang, 2024). Recent studies have further characterized learning dynamics through long-term trends, short-term fluctuations, temporal granularity, and frequency patterns (Hou et al., 2025; Bai et al., 2025a; 2026).

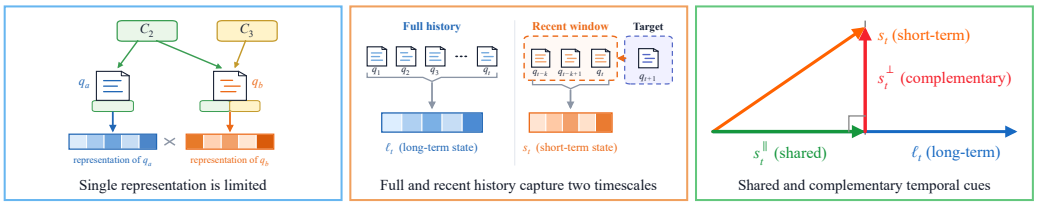
As shown in Figure 1(a), KT predicts a student's response to the next question from question–KC associations, historical outcomes, and interaction order. Figure 1(b) highlights three interrelated modeling challenges. First, the target question requires a precise representation. Methods based mainly on KC identifiers or independent question IDs cannot jointly capture question identity, question–KC relations, multi-KC composition, and difficulty (Sonkar et al., 2020; Chen et al., 2023; Deng et al., 2025). Second, the complete history and target-relevant interactions play distinct temporal roles: the former captures cumulative learning dynamics, whereas the latter provides short-term information for the current prediction. Finally, the resulting long- and short-term states are

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

interdependent. The short-term state may contain both a component aligned with the long-term direction and an orthogonal residual, yet direct concatenation, addition, or holistic gating cannot explicitly distinguish and regulate them.



(a) An example of the knowledge tracing task



(b) Three key challenges in fine-grained knowledge tracing

Figure 1: Knowledge tracing task and core modeling motivation. (a) Given historical question–KC associations, responses, and interaction order, KT predicts the response to the next target question. (b) The three panels illustrate question-level discrimination, separate modeling of the complete-history long-term state and target-relevant short-term state, and decomposition of the short-term state into a long-term-aligned component and an orthogonal residual.

To address these challenges, we propose **DualKT**, a dual-timescale KT model with orthogonal decomposition and target-conditioned fusion. DualKT combines question identity, question–KC graph structure, an aggregated multi-KC representation, and a Rasch-style question-difficulty residual into a structured question representation for interaction encoding, target-relevant retrieval, and state fusion (Rasch, 1993). A Mamba-2 branch encodes the complete learning trajectory into a cumulative long-term state (Dao & Gu, 2024). A hierarchical Transformer queries recent interactions and chunked full-prefix summaries with the target question to obtain a target-relevant short-term state (Vaswani et al., 2017), preserving recent detail while retaining coarse context from the complete visible history. DualKT then projects the short-term state onto the long-term direction, yielding a long-term-aligned component and an orthogonal residual. Target-conditioned fusion routes between long- and short-term information and separately calibrates the orthogonal residual path.

Our contributions are summarized as follows:

- We introduce a dual-timescale state-modeling and target-conditioned fusion framework. DualKT derives a long-term state from the complete trajectory and a target-relevant short-term state from hierarchical memory, decomposes the short-term state into a long-term-aligned component and an orthogonal residual, and lets the target question control their contributions to the current prediction.
- We develop a structured question representation that combines question identity, question–KC graph structure, an aggregated multi-KC representation, and question difficulty, providing a discriminative target condition for long-term encoding, local evidence retrieval, and state fusion.
- We conduct question-level five-fold evaluation on five public datasets. DualKT obtains the highest AUC on four datasets under fixed configurations; after unified global TPE optimization, it ranks first in AUC on three datasets and in accuracy on four. Ablations further show that structured question representation yields the largest and most consistent gain, while the full model outperforms its single-branch variants overall.

2 METHOD

2.1 PROBLEM FORMULATION AND MODEL OVERVIEW

Let $\mathcal{Q}$ and $\mathcal{C}$ denote the sets of questions and knowledge components (KCs), respectively. A learning sequence for student u is $\mathcal{S}^u = \{(q_t, \mathcal{C}_t, r_t)\}_{t=1}^{T}$, where $q_t \in \mathcal{Q}$, $\mathcal{C}_t \subseteq \mathcal{C}$, and $r_t \in \{0, 1\}$ denote the question, its associated KC set, and the response at interaction t. Given the history prefix $\mathcal{H}_t = \{(q_i, \mathcal{C}_i, r_i)\}_{i=1}^{t}$ and the next target question $(q_{t+1}, \mathcal{C}_{t+1})$, KT estimates

$$\hat{p}_{t+1} = P(r_{t+1} = 1 \mid \mathcal{H}_t, q_{t+1}, \mathcal{C}_{t+1}) . \tag{1}$$

We use d for the hidden dimension, $[\cdot;\cdot]$ for concatenation, and $\odot$ for elementwise multiplication. The functions $\sigma(\cdot)$, $\mathrm{LN}(\cdot)$, and $\mathrm{Drop}(\cdot)$ denote sigmoid, layer normalization, and dropout, respectively. Let $\mathcal{C}_q$ be the KC set associated with question q, and let $\pi(q)$ identify its complete KC bundle. Different $\phi_*(\cdot)$ denote independently parameterized feed-forward networks, and $\varepsilon > 0$ is a numerical stability constant.

Figure 2 presents the overall architecture of DualKT. We next describe its components.

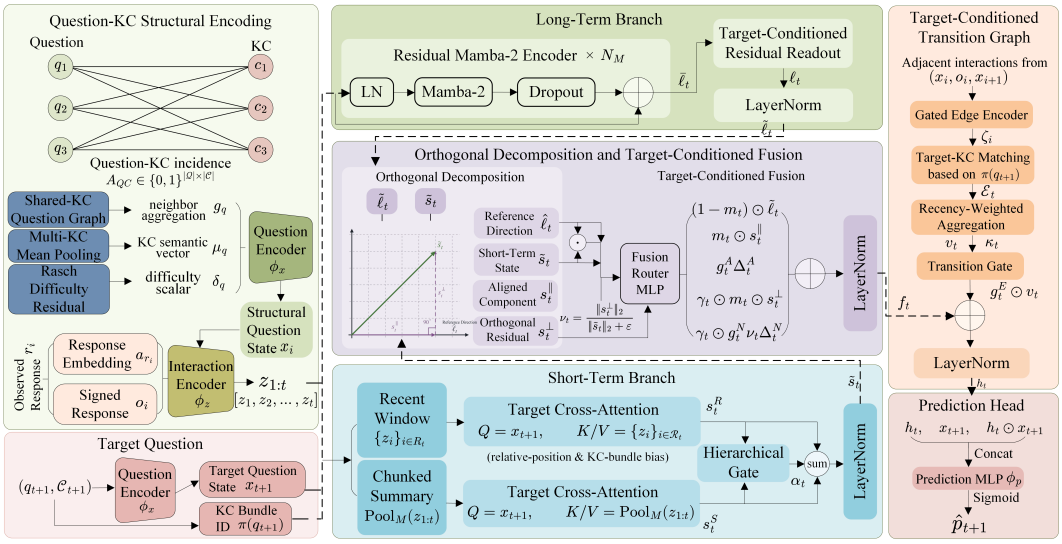


Figure 2: Overview of the DualKT framework. DualKT first encodes question–KC structure and observed responses, then uses the long-term branch to model cumulative history and the short-term branch to retrieve target-relevant information. The short-term state is decomposed relative to the long-term direction for target-conditioned fusion, after which a target-conditioned transition graph refines the fused state for response prediction.

2.2 STRUCTURED QUESTION AND INTERACTION ENCODING

Structured question representation. KC identifiers alone neither exploit question relations induced by shared KCs nor distinguish questions that share KCs but differ in identity or difficulty. Let $A_{QC} \in \{0, 1\}^{|\mathcal{Q}|\times|\mathcal{C}|}$ be the question–KC incidence matrix, define $\overline{A}_{QC} = \mathrm{RowNorm}(A_{QC})$ and $\overline{A}_{CQ} = \mathrm{RowNorm}(A_{QC}^{\top})$, and let E_Q and E_C be learnable question and KC embeddings. One question–KC–question propagation step yields the graph context $g_q = [\overline{A}_{QC}\overline{A}_{CQ}E_Q]_q$, while $\mu_q = |\mathcal{C}_q|^{-1}\sum_{c\in\mathcal{C}_q}[E_C]_c$ represents the mean semantics of the associated KCs.

DualKT further introduces a learnable Rasch-style question-difficulty residual δ_q (Rasch, 1993), yielding the structural representation $\chi_q = \mathrm{LN}([E_Q]_q + \mathrm{Drop}(W_g g_q) + \mathrm{Drop}(W_c\mu_q) + \tanh(\delta_q)\mu_q)$. The final question representation retains the elementwise interaction between KC semantics and this structural representation:

$$x_t = \mathrm{LN}(\phi_x([\mu_{q_t}; \chi_{q_t}; \mu_{q_t} \odot \chi_{q_t}])) . \tag{2}$$

Here, μ_q summarizes the semantics of the associated KCs, whereas χ_q captures question identity, shared-KC structure, and question difficulty. Together, they form the structured question representation x_t used for history encoding, target retrieval, and state fusion (Sonkar et al., 2020; Chen et al., 2023; Deng et al., 2025).

Interaction representation. Let $a_{r_t} \in \mathbb{R}^d$ be a binary response embedding and $o_t = \tanh(W_o(r_t - 0.5) + b_o)$ be a signed response representation. The former encodes the response category, the latter preserves its signed direction, and the product with x_t makes the update question dependent:

$$z_t = \mathrm{LN}(\phi_z([x_t; a_{r_t}; x_t \odot \tanh(a_{r_t}); o_t])) \,. \tag{3}$$

2.3 Dual-Timescale Cognitive State Modeling

Long-term state. The long-term branch tracks cumulative cognitive evolution by applying a residual Mamba-2 encoder (Gu & Dao, 2023; Dao & Gu, 2024) causally to the complete visible prefix $z_{1:t}$. Let $\bar{\ell}_t = [\mathrm{Mamba2}(z_1, \ldots, z_t)]_t$ be its current output. The target question then emphasizes dimensions relevant to the present prediction:

$$\ell_t = \bar{\ell}_t \odot \big(1 + \sigma\big(\phi_L([\bar{\ell}_t; x_{t+1}])\big)\big) \,. \tag{4}$$

Because $1 + \sigma(\cdot) \in (1, 2)^d$, this residual gate emphasizes target-relevant dimensions without discarding accumulated information, yielding the target-conditioned long-term state ℓ_t.

Target-relevant short-term state. The complete history can dilute interactions that are particularly informative for the next question. The short-term branch therefore uses x_{t+1} to query both fine-grained recent memory and chunked full-prefix summaries (Vaswani et al., 2017; Pandey & Karypis, 2019; Ghosh et al., 2020). Let $\mathcal{R}_t$ contain the most recent W_R visible interactions and d_k be the dimension of head k. For $i \in \mathcal{R}_t$, content similarity is combined with a temporal–KC bias $\omega_{t,i}^{(k)}$ that accounts for interaction distance and exact KC bundle matching:

$$\xi_{t,i}^{(k)} = \frac{\left\langle W_Q^{(k)} x_{t+1}, W_K^{(k)} z_i \right\rangle}{\sqrt{d_k}} + \omega_{t,i}^{(k)} \,. \tag{5}$$

Let TCA denote multi-head target cross-attention and let Ψ_t collect these biases. It produces the recent-memory readout $s_t^R = \mathrm{TCA}(x_{t+1}, \{z_i\}_{i \in \mathcal{R}_t}; \Psi_t)$; the exact bias is given in Appendix B.

Recent memory preserves local detail but may miss broader context. DualKT therefore mean-pools non-overlapping chunks of size M and queries the ordered summaries together with a learnable null token $z_\varnothing$, yielding $s_t^S = \mathrm{TCA}(x_{t+1}, [z_\varnothing; \mathrm{Pool}_M(z_{1:t})])$. A target-conditioned gate then combines the two resolutions:

$$s_t = \mathrm{LN}\big(\alpha_t \odot s_t^R + (1 - \alpha_t) \odot s_t^S\big) \,. \tag{6}$$

2.4 Orthogonal Decomposition and Target-Conditioned Fusion

Treating the short-term state as a single block would mix information that follows the long-term tendency with information that departs from it. DualKT therefore uses the long-term direction as a geometric reference. Let $\widetilde{\ell}_t = \mathrm{LN}_L(\ell_t)$, $\widetilde{s}_t = \mathrm{LN}_S(s_t)$, and $\widehat{\ell}_t = \widetilde{\ell}_t / \max(\|\widetilde{\ell}_t\|_2, \varepsilon)$. The short-term state is decomposed as

$$(s_t^{\parallel}, s_t^{\perp}) = \Big(\langle \widetilde{s}_t, \widehat{\ell}_t \rangle \widehat{\ell}_t, \widetilde{s}_t - \langle \widetilde{s}_t, \widehat{\ell}_t \rangle \widehat{\ell}_t\Big) \,. \tag{7}$$

Here, $s_t^{\parallel}$ is the signed projection along the long-term direction, while $s_t^{\perp}$ is the remaining residual; together they reconstruct $\widetilde{s}_t$. When $\|\widetilde{\ell}_t\|_2 \geq \varepsilon$, the residual is orthogonal to $\widehat{\ell}_t$.

The target question next determines how these components should affect the prediction. An elementwise router m_t balances long- and short-term information, while a separate factor γ_t calibrates whether the orthogonal residual is useful for the current target. To condition this calibration on both quantities, DualKT forms $c_t^{\perp} = [s_t^{\perp}; x_{t+1}; s_t^{\perp} \odot x_{t+1}; |s_t^{\perp} - x_{t+1}|]$ and computes

$$\gamma_t = 1 + \tanh\big(W_\gamma c_t^{\perp} + b_\gamma\big) \,. \tag{8}$$

The factor $\gamma_t \in (0,2)^d$ can suppress or amplify the residual path. Specifically, $u_t^A = (1-m_t)\odot\widetilde{\ell}_t + m_t \odot s_t^{\parallel} + g_t^A \Delta_t^A$ collects the long-term and aligned terms, whereas $u_t^N = m_t \odot s_t^{\perp} + g_t^N \nu_t \Delta_t^N$ collects the orthogonal residual terms. Here, ν_t is the relative residual magnitude; Appendix B.3 gives the compact correction terms. Their target-conditioned combination is

$$f_t = \mathrm{LN}\big(u_t^A + \gamma_t \odot u_t^N\big)\,. \tag{9}$$

This design preserves the accumulated state as a reference while allowing target-specific deviations to contribute when informative.

2.5 Transition-Graph-Enhanced Prediction

The fused state summarizes cognitive history but does not explicitly preserve how one question leads to the next. As a lightweight local correction, DualKT builds a target-conditioned transition graph over adjacent interactions. Each edge representation ζ_i encodes its endpoint questions and the observed source response. For the current candidate edge $t \to t+1$, this response is r_t; the unknown r_{t+1} is never used. DualKT retains edges whose destination KC bundle matches the target, $\mathcal{E}_t = \{i \mid 1 \le i \le t, \pi(q_{i+1}) = \pi(q_{t+1})\}$, and aggregates them with recency decay $\rho_e \in (0,1)$:

$$v_t = \mathrm{LN}\left(W_v \frac{\sum_{i\in\mathcal{E}_t} \rho_e^{t-i}\zeta_i}{\sum_{i\in\mathcal{E}_t} \rho_e^{t-i} + \varepsilon}\right). \tag{10}$$

More matched edges provide stronger support. With $N_t^E = |\mathcal{E}_t|$, the coefficient $\kappa_t = N_t^E/(N_t^E + \tau)$ scales the target-conditioned gate $g_t^E = \sigma(\phi_E([f_t; x_{t+1}; v_t]))\kappa_t$, which injects the transition representation residually:

$$h_t = \mathrm{LN}\big(f_t + \mathrm{Drop}(W_{\mathrm{tr}}(g_t^E v_t))\big)\,. \tag{11}$$

The prediction head combines the final state, the target question representation, and their elementwise interaction:

$$\hat{p}_{t+1} = \sigma(\phi_P([h_t; x_{t+1}; h_t \odot x_{t+1}]))\,. \tag{12}$$

All modules are trained end to end with question-level binary cross-entropy. Appendix B.4 gives the edge encoder, valid-position details, and complete objective.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate DualKT on five public datasets: ASSIST2009 and ASSIST2017 from ASSISTments (Feng et al., 2009), NIPS34 from the NeurIPS 2020 Education Challenge (Wang et al., 2020), XES3G5M from an online mathematics learning platform (Liu et al., 2023c), and Algebra2005 from the KDD Cup 2010 Educational Data Mining Challenge (Stamper et al., 2010). Appendix C.1 describes each dataset, and Table 2 reports statistics before sliding-window segmentation.

Baselines. We compare DualKT with 30 representative models: recurrent and memory-based DKT, DKT+, DKT-F, KQN, DKVMN, SKVMN, Deep-IRT, and qDKT (Piech et al., 2015; Yeung & Yeung, 2018; Nagatani et al., 2019; Lee & Yeung, 2019; Zhang et al., 2017; Abdelrahman & Wang, 2019; Yeung, 2019; Sonkar et al., 2020); graph-, attention-, and temporal-based GKT, SAKT, SAINT, AKT, RKT, HawkesKT, ATKT, and DIMKT (Nakagawa et al., 2019; Pandey & Karypis, 2019; Choi et al., 2020; Ghosh et al., 2020; Pandey & Srivastava, 2020; Wang et al., 2021; Guo et al., 2021; Shen et al., 2022); question-enhanced AT-DKT, simpleKT, QIKT, DTransformer, FoLiBiKT, extraKT, DenoiseKT, and FA-KT (Liu et al., 2023a;b; Chen et al., 2023; Yin et al., 2023; Im et al., 2023; Li et al., 2024; Deng et al., 2025; Bai et al., 2026); and multi-timescale, state-space, and uncertainty-aware MTKT, LefoKT, FlucKT, UKT, Mamba4KT, and MCKT (Bai et al., 2025a;b; Hou et al., 2025; Cheng et al., 2025; Cao & Zhang, 2024; Yao & Yu, 2026). Appendix C.2 gives a brief description of every baseline.

Metrics and data splits. AUC is the primary metric and ACC the secondary metric. Absolute AUC and ACC differences are reported in % (i.e., multiplied by 100). Following pyKT (Liu et al., 2022)[1],

[1] https://pykt.org/

we discard sequences with fewer than three interactions, cap histories at 200, and split students into 80% training–validation and 20% held-out test data. Training and tuning use no test information, and models and hyperparameters are selected by validation AUC.

Unified question-level evaluation. KC-sequence models use question-level sliding windows and LF-AVG to aggregate multi-KC predictions, whereas native question-level models directly predict question probabilities. All models share student splits, maximum history length, and final labels while retaining their native inputs. The question graph uses only static question–KC metadata, and data-dependent statistics, including question difficulty, are estimated from the training fold.

Hyperparameters and implementation. Fixed-configuration comparisons use published settings or public-implementation defaults. For recent KT models, unified global tree-structured Parzen estimator (TPE) optimization evaluates at least 200 configurations per model–dataset pair, using validation AUC only. Search spaces cover learning rate, dropout, representation dimension, network depth, and model-specific parameters. All experiments run on an NVIDIA A800 Linux server. DualKT uses 128-dimensional states, two Mamba-2 layers, two Transformer layers, four attention heads, a 64-interaction recent window, 32-interaction summary chunks, and 0.2 dropout. We train with Adam, batch size 64, learning rate 10^{-3}, at most 200 epochs, early-stopping patience 10, and gradient-norm clipping at 5.0.

3.2 Overall Performance

Table 1: Overall fixed-configuration AUC on five datasets. Each entry is the five-fold mean±standard deviation; bold and underlined values denote the best and second-best available results in each column. "–" marks unavailable pairs: ASSIST2009 lacks the real timestamps required by DKT-F, MTKT, and FA-KT; DenoiseKT's explicit shared-KC question graph is infeasible for Algebra2005's 173,113 questions.

Model	ASSIST2009	NIPS34	XES3G5M	ASSIST2017	Algebra2005
DKT	0.7514 ± 0.0011	0.7679 ± 0.0003	0.7832 ± 0.0005	0.7180 ± 0.0018	0.8120 ± 0.0009
DKVMN	0.7451 ± 0.0010	0.7665 ± 0.0006	0.7782 ± 0.0003	0.7133 ± 0.0008	0.8003 ± 0.0011
DKT+	0.7441 ± 0.0005	0.7603 ± 0.0009	0.7741 ± 0.0006	0.7064 ± 0.0007	0.8037 ± 0.0009
DKT-F	–	0.7723 ± 0.0003	0.7931 ± 0.0005	0.7231 ± 0.0021	0.8127 ± 0.0020
KQN	0.7453 ± 0.0006	0.7660 ± 0.0004	0.7760 ± 0.0002	0.7182 ± 0.0020	0.7958 ± 0.0026
SKVMN	0.7347 ± 0.0006	0.7497 ± 0.0004	0.7504 ± 0.0009	0.7075 ± 0.0007	0.7364 ± 0.0025
Deep-IRT	0.7439 ± 0.0008	0.7661 ± 0.0007	0.7775 ± 0.0005	0.7114 ± 0.0008	0.8007 ± 0.0013
qDKT	0.6900 ± 0.0054	0.7966 ± 0.0006	0.8166 ± 0.0002	0.7434 ± 0.0016	0.7445 ± 0.0015
GKT	0.7319 ± 0.0037	0.7572 ± 0.0105	0.7721 ± 0.0007	0.7321 ± 0.0014	0.7895 ± 0.0029
SAKT	0.7315 ± 0.0016	0.7435 ± 0.0007	0.7606 ± 0.0002	0.6551 ± 0.0010	0.7851 ± 0.0011
SAINT	0.6955 ± 0.0014	0.7826 ± 0.0011	0.8019 ± 0.0006	0.6778 ± 0.0056	0.7686 ± 0.0049
AKT	0.7616 ± 0.0014	0.7929 ± 0.0003	0.8098 ± 0.0002	0.7426 ± 0.0009	0.8115 ± 0.0029
RKT	0.7480 ± 0.0052	0.7920 ± 0.0007	0.8147 ± 0.0006	0.7130 ± 0.0006	0.7709 ± 0.0008
ATKT	0.7661 ± 0.0007	0.7706 ± 0.0005	0.7982 ± 0.0009	0.7288 ± 0.0008	0.8061 ± 0.0010
HawkesKT	0.7253 ± 0.0005	0.6459 ± 0.0007	0.6932 ± 0.0048	0.6180 ± 0.0012	0.7336 ± 0.0026
DIMKT	0.7637 ± 0.0018	0.8007 ± 0.0004	0.8207 ± 0.0004	$\underline{0.7914 \pm 0.0009}$	0.8269 ± 0.0016
AT-DKT	0.7497 ± 0.0013	0.7803 ± 0.0003	0.7919 ± 0.0006	0.7480 ± 0.0012	0.8146 ± 0.0015
simpleKT	0.7702 ± 0.0013	0.8024 ± 0.0002	0.8121 ± 0.0003	0.7609 ± 0.0015	0.8149 ± 0.0016
QIKT	0.7800 ± 0.0023	0.8022 ± 0.0005	0.8179 ± 0.0004	0.7774 ± 0.0008	0.8188 ± 0.0006
DTransformer	0.7676 ± 0.0029	0.7974 ± 0.0002	0.8091 ± 0.0016	0.7424 ± 0.0018	0.8059 ± 0.0039
FoLiBiKT	0.7768 ± 0.0023	$\underline{0.8031 \pm 0.0003}$	0.8138 ± 0.0006	0.7755 ± 0.0006	0.8154 ± 0.0009
extraKT	0.7758 ± 0.0013	0.8029 ± 0.0002	0.8165 ± 0.0004	0.7588 ± 0.0025	0.8107 ± 0.0018
Mamba4KT	0.6636 ± 0.0026	0.7687 ± 0.0005	0.7992 ± 0.0004	0.7536 ± 0.0003	0.7452 ± 0.0008
MCKT	0.6807 ± 0.0028	0.7895 ± 0.0004	0.8078 ± 0.0004	0.7585 ± 0.0004	0.7310 ± 0.0009
MTKT	–	0.7994 ± 0.0009	0.8146 ± 0.0003	0.7616 ± 0.0015	0.8258 ± 0.0025
LefoKT	0.7690 ± 0.0016	0.7996 ± 0.0004	0.8146 ± 0.0001	0.7422 ± 0.0018	0.8170 ± 0.0033
FlucKT	0.7771 ± 0.0008	0.8023 ± 0.0004	0.8136 ± 0.0009	0.7665 ± 0.0015	0.8287 ± 0.0020
UKT	0.7827 ± 0.0014	0.7927 ± 0.0014	0.8096 ± 0.0004	0.7634 ± 0.0034	$\underline{0.8305 \pm 0.0017}$
DenoiseKT	$\mathbf{0.7839 \pm 0.0022}$	0.7994 ± 0.0003	$\underline{0.8218 \pm 0.0004}$	0.7715 ± 0.0019	–
FA-KT	–	0.8029 ± 0.0003	0.8198 ± 0.0005	0.7780 ± 0.0010	$\underline{0.8305 \pm 0.0014}$
DualKT (ours)	$\underline{0.7833 \pm 0.0016}$	$\mathbf{0.8043 \pm 0.0001}$	$\mathbf{0.8282 \pm 0.0003}$	$\mathbf{0.7967 \pm 0.0008}$	$\mathbf{0.8389 \pm 0.0006}$

Table 1 shows that DualKT ranks among the top two methods on all five datasets and performs best on NIPS34, XES3G5M, ASSIST2017, and Algebra2005. Relative to the strongest available baseline, it gains 0.12%, 0.64%, 0.53%, and 0.84% on these four datasets, respectively, while trailing DenoiseKT by only 0.06% on ASSIST2009; the mean dataset-wise margin is 0.41%. The 0.53% and 0.84% gains on ASSIST2017 and Algebra2005, the two longest-history datasets, indicate that the advantage persists when longer histories must be summarized.

Table 3 shows that DualKT also ranks first in ACC on ASSIST2009, XES3G5M, and ASSIST2017, exceeding the strongest baseline by 0.27%, 0.24%, and 0.45%, respectively. It ranks second on NIPS34 and Algebra2005, with gaps of only 0.01% and 0.43%. The agreement between AUC and ACC suggests that its gains in probability ranking generally translate into reliable response classification.

3.3 Unified Bayesian Optimization Comparison

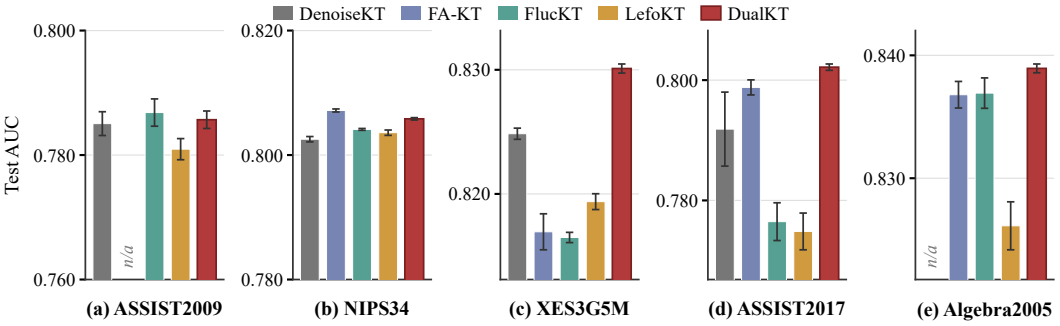


Figure 3: Test AUC under unified global TPE optimization. Bars and error bars show five-fold means $\pm$ one standard deviation.

Figure 3 compares five models under the same global TPE search budget. DualKT trails FlucKT on ASSIST2009 and FA-KT on NIPS34 by only 0.12% and 0.13%, respectively. On XES3G5M, ASSIST2017, and Algebra2005, it exceeds the strongest competitor by 0.53%, 0.34%, and 0.20%, respectively. This ranking agrees with the fixed-configuration comparison, indicating that DualKT's advantage is not attributable to a larger tuning budget. The corresponding ACC results in Figure 8 show that DualKT ranks first on four datasets and second on NIPS34.

3.4 Ablation Study

Figure 4 compares seven core variants against the full model. Each cell reports the AUC or ACC change relative to full DualKT.

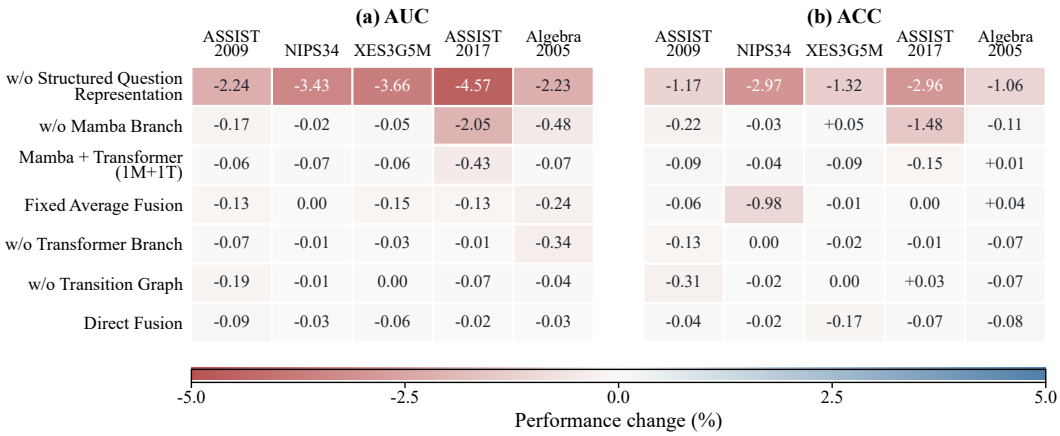


Figure 4: AUC and ACC changes relative to full DualKT across seven ablations and five datasets. Each cell reports $100 \times (\text{variant} - \text{full})$ in %; negative values indicate degradation.

In particular, *w/o Structured Question Representation* removes question identity, message passing over the shared-KC question graph, the Rasch-style question-difficulty residual, and question–KC

fusion, leaving only the KC-bundle representation for each interaction. *Direct Fusion* replaces the proposed orthogonal fusion with a linear projection of concatenated long- and short-term states.

Removing the structured question representation reduces mean AUC and ACC by 3.23% and 1.90%, respectively, with both metrics degrading on every dataset. This result identifies it as the most consequential component of DualKT.

The full two-layer Mamba plus two-layer Transformer architecture outperforms the Mamba-only, Transformer-only, and shallower 1M+1T variants in AUC on every dataset, while also achieving higher macro-average ACC than all three variants. This pattern supports the complementarity of long-term state propagation, target-relevant short-term retrieval, and sufficient encoder depth.

Ablation sensitivity is dataset specific: removing Mamba reduces AUC by 2.05% on ASSIST2017 versus at most 0.48% elsewhere, while fixed fusion preserves NIPS34 AUC but lowers ACC by 0.98%.

Fixed averaging and Direct Fusion reduce macro-average AUC/ACC by 0.13/0.20% and 0.05/0.08%, respectively; Direct Fusion is worse on all five datasets, supporting learned geometry-aware fusion. Removing the transition graph has only small effects.

3.5 Long-Sequence Robustness

To evaluate robustness to history length, we vary the maximum sequence length from 200 to 1,000. At each length, all models are trained from scratch using identical data splits, while all other experimental settings are held fixed.

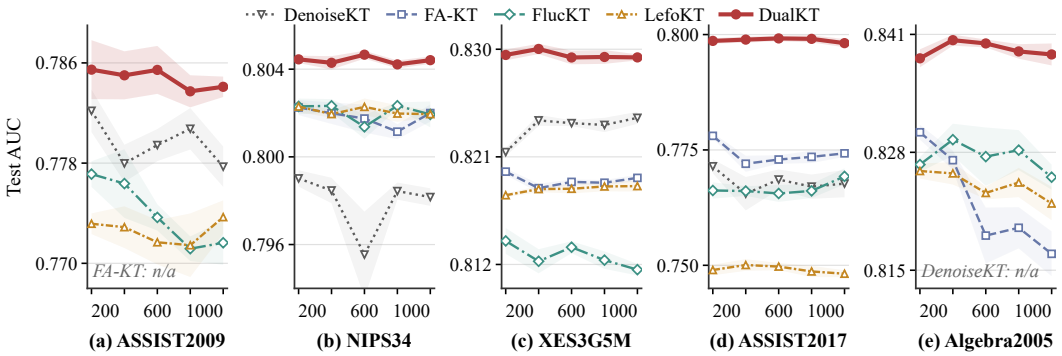


Figure 5: Long-sequence robustness across five datasets. Curves and bands show five-fold mean test AUC $\pm$ one standard deviation as the maximum sequence length increases from 200 to 1,000.

Figure 5 compares five models as the maximum sequence length increases from 200 to 1,000. DualKT achieves the highest AUC in all 25 supported dataset–length comparisons, while its five-dataset macro-average changes only from 0.8113 to 0.8110 (a 0.03% decrease). At length 1,000, its margins over the strongest supported baseline are 0.64%, 0.24%, 0.51%, 2.39%, and 1.35% on ASSIST2009, NIPS34, XES3G5M, ASSIST2017, and Algebra2005, respectively. It likewise ranks first in all 25 supported ACC comparisons in Figure 9.

Across the full sweep, DualKT's within-dataset AUC range is only 0.04%–0.20%, and every fold standard deviation is at most 0.24%. Thus, its lead is not produced by a single favorable truncation length and remains stable relative to cross-fold variation.

3.6 Dynamic Routing and State Geometry

Figure 6 aligns prediction confidence and correctness with the mean short-term routing weight and long–short cosine similarity over the same 100-step sequence. The routing weight changes frequently, including between adjacent predictions, rather than following a fixed or monotonic schedule; meanwhile, state directions alternate among opposed, approximately orthogonal, and aligned regimes. The two tracks are not synchronized: high short-term weights can coincide with low

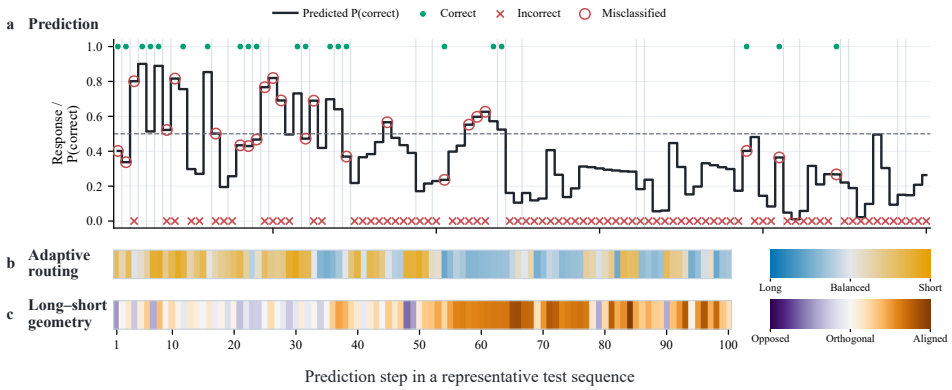


Figure 6: Dynamic routing on an ASSIST2017 sequence: (a) predictions and outcomes, (b) mean short-term routing weight, and (c) long–short cosine similarity; continuous outputs are unsmoothed.

or negative cosine similarity, while aligned states do not always receive the maximum weight. This behavior indicates that target-conditioned fusion selectively retains complementary evidence rather than using state similarity itself as the routing score. Figure 7 provides aggregate evidence for the same pattern.

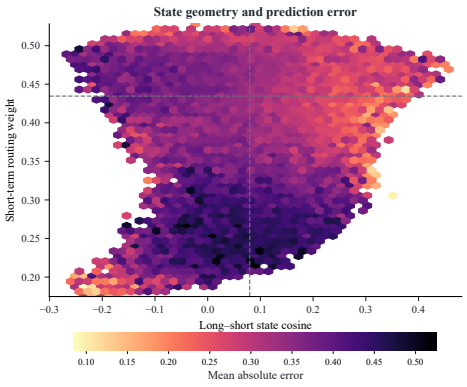


Figure 7: State geometry on ASSIST2017 fold 0 (n = 177,767; $\geq$ 15 predictions per hexagon). Color denotes mean absolute error; dashed lines mark medians.

Across 177,767 fold-0 predictions, the median state cosine and short-term routing weight are 0.080 and 0.435. The four median-defined regions contain 41,690–47,193 predictions each; their mean absolute errors decrease from 0.372 and 0.355 to 0.334 and 0.300 as similarity and routing weight increase, yielding a 19.4% endpoint reduction. The broad weight range at similar cosine values, and conversely similar weights under opposed and aligned states, shows that routing is not a deterministic transform of cosine similarity. Nevertheless, lower error shifts toward jointly stronger alignment and short-term routing, while high weights across the similarity range indicate that nonredundant short-term evidence can be retained. Together with Figure 6 and the fixed-fusion ablation, this pattern further supports state-dependent long–short coordination and the retention of complementary short-term evidence.

4 Conclusion

We presented DualKT, a dual-timescale knowledge tracing model that distinguishes long-term cognitive evolution from target-relevant short-term variation. DualKT constructs a structured question representation, encodes the complete trajectory with Mamba-2, and retrieves a target-relevant short-term state with a hierarchical Transformer. It decomposes the short-term state into a long-term-aligned component and an orthogonal residual, then combines them through target-conditioned fusion. Question-level five-fold experiments on five public datasets show that DualKT ranks among the top two methods in AUC on every dataset and first on four under fixed configurations. Ablations identify the structured question representation as the main source of improvement, with further gains from the dual-timescale branches and learned fusion. State and routing analyses show that DualKT adjusts long- and short-term contributions rather than using a fixed mixture. These results demonstrate the effectiveness of explicitly separating and dynamically integrating long- and short-term information for knowledge tracing.

AI use statement. Generative AI assisted literature organization, translation, and language editing; the authors verified all scientific content, and no experimental data were AI-generated.

REFERENCES

Ghodai Abdelrahman and Qing Wang. Knowledge tracing with sequential key-value memory networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 175–184, 2019. doi: 10.1145/3331184.3331195.

Youheng Bai, Xueyi Li, Zitao Liu, Yaying Huang, Teng Guo, Mingliang Hou, and Weiqi Luo. Learning multi-granularity temporal characteristics for attention-based knowledge tracing. *Neurocomputing*, 655:131338, 2025a. doi: 10.1016/j.neucom.2025.131338.

Youheng Bai, Xueyi Li, Zitao Liu, Yaying Huang, Mi Tian, and Weiqi Luo. Rethinking and improving student learning and forgetting processes for attention based knowledge tracing models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27822–27830, 2025b. doi: 10.1609/aaai.v39i27.34998.

Youheng Bai, Mingliang Hou, Teng Guo, Zitao Liu, and Weiqi Luo. A frequency-aware mixture of heterogeneous experts framework for knowledge tracing. In *Proceedings of the ACM Web Conference 2026*, pp. 7080–7090, 2026. doi: 10.1145/3774904.3792272.

Yang Cao and Wei Zhang. Mamba4KT: An efficient and effective mamba-based knowledge tracing model. *arXiv preprint arXiv:2405.16542*, 2024. URL https://arxiv.org/abs/2405.16542.

Jiahao Chen, Zitao Liu, Shuyan Huang, Qiongqiong Liu, and Weiqi Luo. Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14196–14204, 2023. doi: 10.1609/aaai.v37i12.26661.

Weihua Cheng, Hanwen Du, Chunxiao Li, Ersheng Ni, Liangdi Tan, Tianqi Xu, and Yongxin Ni. Uncertainty-aware knowledge tracing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27905–27913, 2025. doi: 10.1609/aaai.v39i27.35007.

Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, and Jaewe Heo. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the Seventh ACM Conference on Learning at Scale*, pp. 341–344, 2020. doi: 10.1145/3386527.3405945.

Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995. doi: 10.1007/BF01099821.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Jiwei Deng, Youheng Bai, Mingliang Hou, Teng Guo, Zitao Liu, and Weiqi Luo. Denoised attention and question-augmented representations for knowledge tracing. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9619–9627, 2025. doi: 10.24963/ijcai.2025/1069.

Mingyu Feng, Neil T. Heffernan, and Kenneth R. Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19 (3):243–266, 2009. doi: 10.1007/s11257-009-9063-7.

Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2330–2339, 2020. doi: 10.1145/3394486.3403282.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL https://arxiv.org/abs/2312.00752.

Teng Guo, Yu Qin, Yubin Xia, Mingliang Hou, Zitao Liu, Feng Xia, and Weiqi Luo. Enhancing knowledge tracing through decoupling cognitive pattern from error-prone data. In *Proceedings of the ACM on Web Conference 2025*, pp. 5108–5116, 2025. doi: 10.1145/3696410.3714486.

Xiaopeng Guo, Zhijie Huang, Jie Gao, Mingyu Shang, Maojing Shu, and Jun Sun. Enhancing knowledge tracing via adversarial training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 367–375, 2021. doi: 10.1145/3474085.3475554.

Mingliang Hou, Xueyi Li, Teng Guo, Zitao Liu, Mi Tian, Renqiang Luo, and Weiqi Luo. Cognitive fluctuations enhanced attention network for knowledge tracing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):14265–14273, 2025. doi: 10.1609/aaai.v39i13.33562.

Yoonjin Im, Eunseong Choi, Heejin Kook, and Jongwuk Lee. Forgetting-aware linear bias for attentive knowledge tracing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3958–3962, 2023. doi: 10.1145/3583780.3615191.

Jinseok Lee and Dit-Yan Yeung. Knowledge query network for knowledge tracing: How knowledge interacts with skills. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge*, pp. 491–500, 2019. doi: 10.1145/3303772.3303786.

Xueyi Li, Youheng Bai, Teng Guo, Ying Zheng, Mingliang Hou, Bojun Zhan, Yaying Huang, Zitao Liu, Boyu Gao, and Weiqi Luo. Extending context window of attention based knowledge tracing models via length extrapolation. In *ECAI 2024*, Frontiers in Artificial Intelligence and Applications. IOS Press, 2024. doi: 10.3233/FAIA240651.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Jiliang Tang, and Weiqi Luo. pyKT: A python library to benchmark deep learning based knowledge tracing models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18542–18555, 2022. URL https://arxiv.org/abs/2206.11460.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Boyu Gao, Weiqi Luo, and Jian Weng. Enhancing deep knowledge tracing with auxiliary tasks. In *Proceedings of the ACM Web Conference 2023*, pp. 4178–4187, 2023a. doi: 10.1145/3543507.3583866.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simpleKT: A simple but tough-to-beat baseline for knowledge tracing. *arXiv preprint arXiv:2302.06881*, 2023b. URL https://arxiv.org/abs/2302.06881.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. XES3G5M: A knowledge tracing benchmark dataset with auxiliary information. In *Advances in Neural Information Processing Systems 36*, pp. 32958–32970, 2023c. doi: 10.52202/075280-1429.

Koki Nagatani, Qian Zhang, Masahiro Sato, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. Augmenting knowledge tracing by considering forgetting behavior. In *The World Wide Web Conference*, pp. 3101–3107, 2019. doi: 10.1145/3308558.3313565.

Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 156–163, 2019. doi: 10.1145/3350546.3352513.

Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. In *Proceedings of the 12th International Conference on Educational Data Mining*, pp. 384–389, 2019. URL https://educationaldatamining.org/EDM2019/proceedings/.

Shalini Pandey and Jaideep Srivastava. RKT: Relation-aware self-attention for knowledge tracing. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 1205–1214, 2020. doi: 10.1145/3340531.3411994.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, volume 28, pp. 505–513, 2015.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, Chicago, IL, 1993. Reprint of the work originally published in 1960.

Shuanghong Shen, Zhenya Huang, Qi Liu, Yu Su, Shijin Wang, and Enhong Chen. Assessing student's dynamic knowledge state by exploring the question difficulty effect. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 427–437, 2022. doi: 10.1145/3477495.3531939.

Shashank Sonkar, Andrew E. Waters, Andrew S. Lan, Phillip J. Grimaldi, and Richard G. Baraniuk. qDKT: Question-centric deep knowledge tracing. In *Proceedings of the 13th International Conference on Educational Data Mining*, 2020. URL https://arxiv.org/abs/2005.12442.

John Stamper, Alexandru Niculescu-Mizil, Steven Ritter, Geoffrey J. Gordon, and Kenneth R. Koedinger. Algebra I 2005–2006: Challenge data set from KDD Cup 2010 educational data mining challenge. PSLC DataShop, 2010. URL https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Chenyang Wang, Weizhi Ma, Min Zhang, Chuancheng Lv, Fengyuan Wan, Huijie Lin, Taoran Tang, Yiqun Liu, and Shaoping Ma. Temporal cross-effects in knowledge tracing. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 517–525, 2021. doi: 10.1145/3437963.3441802.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E. Turner, Richard G. Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. Diagnostic questions: The NeurIPS 2020 education challenge. *arXiv preprint arXiv:2007.12061*, 2020. URL https://arxiv.org/abs/2007.12061.

Yumo Yao and Hai-Tao Yu. MCKT: A parallel mamba-based model for content-aware knowledge tracing. *IEEE Access*, 14:70510–70529, 2026. doi: 10.1109/ACCESS.2026.3688703.

Chun-Kit Yeung. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*, 2019. URL https://arxiv.org/abs/1904.11738.

Chun-Kit Yeung and Dit-Yan Yeung. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pp. 1–10, 2018. doi: 10.1145/3231644.3231647.

Yu Yin, Le Dai, Zhenya Huang, Shuanghong Shen, Fei Wang, Qi Liu, Enhong Chen, and Xin Li. Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer. In *Proceedings of the ACM Web Conference 2023*, pp. 855–864, 2023. doi: 10.1145/3543507.3583255.

Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 765–774, 2017. doi: 10.1145/3038912.3052580.

A Related Work

A.1 Sequential Knowledge Tracing

Knowledge tracing estimates a student's evolving latent knowledge state from historical learning interactions. Bayesian Knowledge Tracing models mastery through transitions between latent variables, whereas Deep Knowledge Tracing encodes interaction sequences as continuous states with recurrent neural networks (Corbett & Anderson, 1995; Piech et al., 2015). Subsequent work introduced structured memory and representation mechanisms: DKVMN separates KCs from their mastery values in a dynamic key–value memory, KQN maps learner states and KC representations into a shared space, and SKVMN combines sequential memory with attentive retrieval (Zhang et al., 2017; Lee & Yeung, 2019; Abdelrahman & Wang, 2019). Selective state-space models provide an efficient alternative for long-sequence recurrence; Mamba4KT adapts Mamba to KT, while MCKT uses parallel Mamba streams to combine question content with perceived difficulty (Gu & Dao, 2023; Dao & Gu, 2024; Cao & Zhang, 2024; Yao & Yu, 2026). These models are effective at updating a state along the complete history, but a single recurrent state or memory readout must often represent both persistent trends and rapidly changing local evidence without explicitly distinguishing their temporal roles.

A.2 Attention-Based Knowledge Tracing

Attention-based KT allows a target question to retrieve relevant historical interactions directly. SAKT uses a candidate exercise as the query for historical relevance, SAINT models questions and responses with an encoder–decoder architecture, and AKT combines context-aware representations with monotonic attention to capture historical dependence and forgetting (Pandey & Karypis, 2019; Choi et al., 2020; Ghosh et al., 2020). SimpleKT, DTransformer, and DenoiseKT extend this paradigm through architectural simplification, stable knowledge-state estimation, and question-enhanced attention denoising, respectively (Liu et al., 2023b; Yin et al., 2023; Deng et al., 2025). To accommodate non-uniform learning dynamics, FoLiBiKT introduces forgetting-aware attention biases, FlucKT jointly models long-term trends and short-term fluctuations, MTKT learns temporal features at multiple granularities, and FA-KT coordinates heterogeneous experts with frequency-aware routing (Im et al., 2023; Hou et al., 2025; Bai et al., 2025a; 2026). These studies establish the importance of target-conditioned retrieval and multiscale inductive biases. Their cross-scale information, however, is primarily combined through attention scores, temporal features, or expert weights. Such direct fusion can entangle redundant and incremental information in the short-term state, potentially amplifying content that overlaps with the long-term state while weakening the contribution of local changes to the current prediction.

A.3 Graph- and Relation-Aware Knowledge Tracing

Structural relations between questions and KCs provide useful priors for fine-grained cognitive diagnosis. GKT propagates KC states over a knowledge graph, RKT incorporates question relations directly into historical attention, and Deep-IRT and DIMKT characterize student ability, question difficulty, and their interactions through item response theory or explicit difficulty factors (Nakagawa et al., 2019; Pandey & Srivastava, 2020; Rasch, 1993; Yeung, 2019; Shen et al., 2022). qDKT, QIKT, and DenoiseKT further demonstrate that question-level states, joint question–KC prediction, and question-enhanced representations can alleviate the information loss caused by relying on KC labels alone (Sonkar et al., 2020; Chen et al., 2023; Deng et al., 2025). RobustKT also disentangles stable cognitive patterns from stochastic factors in noisy learning records and combines them adaptively, but its objective is robustness to corrupted observations rather than geometric decomposition across temporal scales (Guo et al., 2025). Building on these findings, DualKT uses a question–KC graph to construct a structured question representation that conditions long-term state formation, short-term state retrieval, and subsequent fusion; the target question further determines how the long-term state, long-term-aligned component, and orthogonal residual enter the prediction.

B METHOD DETAILS

B.1 RESIDUAL MAMBA-2 ENCODER

For $Z = [z_1, \ldots, z_T]$, the long-term encoder with N_M state-space layers computes

$$\begin{aligned} Y^{(0)} &= Z, \\ Y^{(n)} &= Y^{(n-1)} + \mathrm{Drop}\Big(\mathrm{Mamba2}^{(n)}\left(\mathrm{LN}^{(n)}(Y^{(n-1)})\right)\Big), \quad n = 1, \ldots, N_M. \end{aligned} \tag{13}$$

The normalized final-layer output at position t gives $\bar{\ell}_t$, which is used to construct the target-conditioned long-term state in Equation 4.

B.2 HIERARCHICAL-MEMORY CONSTRUCTION

At prediction step $t+1$, the recent window and the j-th full-prefix chunk are

$$\begin{aligned} \mathcal{R}_t &= \{i \mid \max(1, t - W_R + 1) \le i \le t\}, \\ \mathcal{I}_{t,j} &= \{i \mid (j-1)M + 1 \le i \le \min(jM, t)\}, \qquad J_t = \left\lceil \frac{t}{M} \right\rceil. \end{aligned} \tag{14}$$

Each chunk contains only visible events and is summarized by

$$\bar{z}_{t,j} = \frac{1}{|\mathcal{I}_{t,j}|} \sum_{i \in \mathcal{I}_{t,j}} z_i. \tag{15}$$

Thus, $\mathrm{Pool}_M(z_{1:t})$ used by the coarse full-prefix readout is the ordered sequence $[\bar{z}_{t,1}; \ldots; \bar{z}_{t,J_t}]$. This memory covers the complete causal prefix and can overlap temporally with the recent window; the two paths differ in resolution rather than using disjoint intervals. The hierarchical gate used in Equation 6 is

$$\alpha_t = \sigma\big(\phi_H([s_t^R; s_t^S; x_{t+1}])\big). \tag{16}$$

The temporal–KC bias in the recent-memory attention score is $\omega_{t,i}^{(k)} = \beta_{t-i}^{(k)} + \lambda_b \rho_b^{t-i} \mathbb{I}[\pi(q_i) = \pi(q_{t+1})]$, where $\beta_{t-i}^{(k)}$ is a relative-position bias, and λ_b and $\rho_b \in (0, 1)$ control the strength and distance decay of the KC bundle matching term.

B.3 FUSION ROUTER, RESIDUAL CALIBRATION, AND CORRECTION HEADS

The independently normalized states are

$$(\widetilde{\ell}_t, \widetilde{s}_t) = (\mathrm{LN}_L(\ell_t), \mathrm{LN}_S(s_t)). \tag{17}$$

The relative magnitude of the orthogonal residual is

$$\nu_t = \frac{\|s_t^{\perp}\|_2}{\|\widetilde{s}_t\|_2 + \varepsilon}. \tag{18}$$

State geometry and the target representation are summarized by

$$c_t^F = \phi_F\Big([\widetilde{\ell}_t; \widetilde{s}_t; x_{t+1}; \widetilde{\ell}_t \odot \widetilde{s}_t; s_t^{\perp}; \nu_t]\Big). \tag{19}$$

The elementwise short-term router and the two scalar correction gates are

$$\begin{aligned} m_t &= \sigma(W_m c_t^F + b_m), \\ [g_t^A, g_t^N] &= \sigma(W_g c_t^F + b_g). \end{aligned} \tag{20}$$

The two correction heads use different state information:

$$\begin{aligned} \Delta_t^A &= \phi_A\Big([\widetilde{\ell}_t \odot \widetilde{s}_t; x_{t+1}]\Big), \\ \Delta_t^N &= \phi_N\big([s_t^{\perp}; s_t^{\perp} \odot x_{t+1}]\big). \end{aligned} \tag{21}$$

Using the residual–target context defined in the main text, its elementwise calibration is

$$\gamma_t = 1 + \tanh\big(W_\gamma c_t^{\perp} + b_\gamma\big)\,. \tag{22}$$

For the compact notation in Equation 9, the aligned and residual groups are

$$\begin{aligned} u_t^A &= (1 - m_t) \odot \widetilde{\ell}_t + m_t \odot s_t^{\parallel} + g_t^A \Delta_t^A, \\ u_t^N &= m_t \odot s_t^{\perp} + g_t^N \nu_t \Delta_t^N. \end{aligned} \tag{23}$$

Thus, m_t is the formal model's only elementwise long–short router, whereas γ_t calibrates only the orthogonal residual path.

B.4 Target-Conditioned Transition Graph and Training Objective

For an adjacent edge $i \to i+1$, DualKT computes

$$\begin{aligned} \widetilde{\zeta}_i &= \phi_{\text{edge}}([x_i; x_{i+1}; o_i; x_i \odot x_{i+1}])\,, \\ \zeta_i &= \sigma(W_{\text{edge}} \widetilde{\zeta}_i + b_{\text{edge}}) \odot \widetilde{\zeta}_i, \end{aligned} \tag{24}$$

where $o_i = \tanh(W_o(r_i - 0.5) + b_o)$. For $i = t$, the destination is the known target question and the response remains the observed source outcome r_t; the construction never uses the unknown r_{t+1}. The target-matched edge set and its size are

$$\begin{aligned} \mathcal{E}_t &= \{i \mid 1 \le i \le t,\ \pi(q_{i+1}) = \pi(q_{t+1})\}, \\ N_t^E &= |\mathcal{E}_t|. \end{aligned} \tag{25}$$

Consequently, the candidate edge $t \to t+1$ belongs to $\mathcal{E}_t$ at every valid prediction position, so $N_t^E \ge 1$; the remaining elements are historical transition edges entering the same target KC bundle. The support coefficient and target-conditioned transition gate are

$$\begin{aligned} \kappa_t &= \frac{N_t^E}{N_t^E + \tau}, \\ g_t^E &= \sigma(\phi_E([f_t; x_{t+1}; v_t]))\,\kappa_t. \end{aligned} \tag{26}$$

The gated transition-graph representation is injected residually:

$$h_t = \mathrm{LN}\big(f_t + \mathrm{Drop}(W_{\text{tr}}(g_t^E v_t))\big)\,. \tag{27}$$

For the valid minibatch positions Ω, all modules are optimized jointly by

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\Omega|} \sum_{(u,t)\in\Omega} \big[r_{t+1}^u \log \hat{p}_{t+1}^u + (1 - r_{t+1}^u) \log(1 - \hat{p}_{t+1}^u)\big]\,. \tag{28}$$

C Experimental Details

Anonymous code. The implementation is available at `https://anonymous.4open.science/r/dualkt-anonymous-source-6F32/`.

C.1 Dataset Descriptions

ASSIST2009. ASSIST2009 contains student–question interactions collected by the ASSISTments tutoring platform during the 2009–2010 school year (Feng et al., 2009). It has the shortest average history among our datasets, but its 17,737 questions and multi-KC annotations provide a challenging test of question-level discrimination.

NIPS34. NIPS34 was released for the NeurIPS 2020 Education Challenge and contains large-scale diagnostic question–response logs (Wang et al., 2020). Its 1.38 million interactions are concentrated over 948 questions and 57 KCs, yielding repeated evidence for estimating student states.

XES3G5M. XES3G5M was collected from a real-world online mathematics learning platform (Liu et al., 2023c). It is the largest dataset in our evaluation by both students and interactions, and its 865 KCs and multi-KC questions test scalability under rich concept structure.

ASSIST2017. ASSIST2017 is another ASSISTments benchmark with relatively few students but long interaction histories. In the processed data, each question is associated with a single KC, allowing evaluation under long sequences without multi-KC ambiguity.

Algebra2005. Algebra2005 originates from the KDD Cup 2010 Educational Data Mining Challenge (Stamper et al., 2010). It combines the longest average histories with a very large question vocabulary and multi-KC questions, making it particularly demanding for both long-range state modeling and question-level generalization.

Table 2 summarizes the dataset statistics before sliding-window segmentation. Average sequence length is computed as the number of interactions divided by the number of students.

Table 2: Statistics of the five KT datasets before sliding-window segmentation. The table reports numbers of students, interactions, unique questions, and KCs, together with average interactions per student (interactions divided by students) and the maximum number of KCs assigned to one question.

Dataset	Students	Interactions	Questions	KCs	Avg. interactions per student	Max KCs per question
ASSIST2009	4,217	346,860	17,737	123	82.3	4
NIPS34	4,918	1,382,727	948	57	281.2	2
XES3G5M	18,066	5,549,635	7,652	865	307.2	6
ASSIST2017	1,709	942,816	3,162	102	551.7	1
Algebra2005	574	607,021	173,113	112	1,057.5	7

C.2 Baseline Descriptions

We compare DualKT with 30 representative KT methods spanning recurrent, memory-based, graph-based, attention-based, question-enhanced, multi-timescale, state-space, and uncertainty-aware designs.

C.2.1 Recurrent and Memory-Based Models

DKT. Deep Knowledge Tracing uses a recurrent neural network to encode question–response interactions into a continuous student state for next-response prediction (Piech et al., 2015).

DKT+. DKT+ augments DKT with prediction-consistency regularization to reduce reconstruction inconsistency and excessive temporal oscillation in its outputs (Yeung & Yeung, 2018).

DKT-F. DKT-F extends recurrent knowledge tracing with forgetting-related temporal features, allowing state updates to account for elapsed time and prior practice (Nagatani et al., 2019).

KQN. The Knowledge Query Network maps the recurrent learner state and the queried skill into a shared space and predicts performance from their interaction (Lee & Yeung, 2019).

DKVMN. DKVMN represents KCs with static memory keys and tracks their evolving mastery with dynamic value slots that are read and updated after each interaction (Zhang et al., 2017).

SKVMN. SKVMN adds sequential dependency modeling to key–value memory retrieval so that successive memory reads preserve temporal structure across learning events (Abdelrahman & Wang, 2019).

Deep-IRT. Deep-IRT combines a DKVMN backbone with item-response-theory variables to expose interpretable student ability and item difficulty alongside predictions (Yeung, 2019).

qDKT. qDKT predicts outcomes at the individual-question level and regularizes related questions through a graph over questions nested under the same skill (Sonkar et al., 2020).

C.2.2 Graph, Attention, and Temporal Models

GKT. Graph-based Knowledge Tracing represents KCs as graph nodes and propagates state changes through their learned or predefined relations with graph neural networks (Nakagawa et al., 2019).

SAKT. SAKT uses the current exercise as a query over past question–response interactions, retrieving history through self-attention rather than recurrent state updates (Pandey & Karypis, 2019).

SAINT. SAINT adopts a Transformer encoder–decoder architecture that separates exercise encoding from response-conditioned decoding for knowledge tracing (Choi et al., 2020).

AKT. AKT combines context-aware question representations with monotonic attention whose decay bias emphasizes educationally relevant and temporally proximate interactions (Ghosh et al., 2020).

RKT. RKT augments self-attention with relation coefficients and forgetting effects so that retrieval reflects both exercise relatedness and temporal distance (Pandey & Srivastava, 2020).

HawkesKT. HawkesKT models learning and forgetting as temporal cross-effects in a multivariate Hawkes process, with interactions exciting or decaying future performance across KCs (Wang et al., 2021).

ATKT. ATKT applies adversarial perturbations during training to improve the robustness of attention-based student-state representations (Guo et al., 2021).

DIMKT. DIMKT explicitly incorporates question and KC difficulty into the interaction representation and matches difficulty against the student's evolving knowledge state (Shen et al., 2022).

C.2.3 Question-Enhanced and Recent Attention Models

AT-DKT. AT-DKT supplements next-response prediction with auxiliary learning objectives that regularize the recurrent representation and preserve educational signals (Liu et al., 2023a).

simpleKT. simpleKT is a deliberately compact attention-based baseline that combines question, KC, and difficulty-aware representations with a standard Transformer encoder (Liu et al., 2023b).

QIKT. QIKT learns question-centric knowledge-acquisition and problem-solving representations and uses an IRT-inspired prediction layer to improve interpretability (Chen et al., 2023).

DTransformer. DTransformer uses a diagnostic Transformer to estimate concept-specific knowledge states and stabilizes them with contrastive learning (Yin et al., 2023).

FoLiBiKT. FoLiBiKT introduces a forgetting-aware linear attention bias that directly encodes recency into attentive retrieval (Im et al., 2023).

extraKT. extraKT redesigns attention biases for length extrapolation so that models trained on bounded windows can generalize to longer interaction sequences (Li et al., 2024).

DenoiseKT. DenoiseKT enriches question representations with difficulty and graph structure and suppresses cognitively irrelevant history through a denoised attention mechanism (Deng et al., 2025).

FA-KT. FA-KT combines self-attention, Mamba, CNN, and LSTM experts and routes among them according to the frequency characteristics of each student sequence (Bai et al., 2026).

C.2.4 MULTI-TIMESCALE, STATE-SPACE, AND UNCERTAINTY-AWARE MODELS

MTKT. MTKT learns temporal characteristics at multiple granularities and integrates them to capture learning dynamics beyond a single temporal resolution (Bai et al., 2025a).

LefoKT. LefoKT uses relative-forgetting attention to separate forgetting patterns from question relevance and to improve extrapolation to longer sequences (Bai et al., 2025b).

FlucKT. FlucKT separates and reweights long-term trends and short-term fluctuations with a causal-convolution decomposition layer and a kernelized attention bias (Hou et al., 2025).

UKT. UKT represents knowledge states as stochastic distributions, applies Wasserstein self-attention, and uses uncertainty-aware contrastive learning for robust prediction (Cheng et al., 2025).

Mamba4KT. Mamba4KT adapts the selective state-space architecture to knowledge tracing for efficient long-sequence training and inference (Cao & Zhang, 2024).

MCKT. MCKT uses parallel Mamba streams to model exercise-content semantics and perceived difficulty, together with adaptive enhancement for short learning histories (Yao & Yu, 2026).

C.3 EVALUATION PROTOCOL

Preprocessing and splits. Following the pyKT protocol (Liu et al., 2022), we remove student sequences with fewer than three interactions and cap each model input at 200 interactions. Students are split into 80% training–validation data and a disjoint 20% held-out test set; five-fold cross-validation is performed only within the training–validation portion.

Question-level alignment and leakage control. KC-sequence models are evaluated through question-level sliding windows, with LF-AVG aggregating the multiple KC predictions associated with one question, whereas native question-level models directly output one probability. All methods use the same student splits, maximum history length, and final question labels. The question graph is constructed only from static question–KC metadata, fold-dependent statistics use the corresponding training fold, and every prediction accesses only its visible prefix.

Metrics and checkpointing. AUC is the primary metric because it evaluates ranking independently of a classification threshold, while ACC provides a complementary thresholded measure. Checkpoints are selected exclusively by validation AUC, and all reported entries are the mean and standard deviation over five independently trained test evaluations.

C.4 FIXED-CONFIGURATION ACC RESULTS

Table 3 complements the primary AUC comparison with classification accuracy under the same fixed configurations. DualKT ranks first on ASSIST2009, XES3G5M, and ASSIST2017, with margins of 0.27%, 0.24%, and 0.45% over the strongest baselines. It ranks second on NIPS34 and Algebra2005, trailing by 0.01% and 0.43%, respectively.

Table 3: Overall fixed-configuration ACC on five datasets. Each entry is the five-fold mean±standard deviation; bold and underlined values denote the best and second-best available results in each column. "–" marks unavailable pairs.

Model	ASSIST2009	NIPS34	XES3G5M	ASSIST2017	Algebra2005
DKT	0.7227 ± 0.0014	0.7027 ± 0.0005	0.8166 ± 0.0004	0.6857 ± 0.0016	0.8085 ± 0.0007
DKVMN	0.7197 ± 0.0005	0.7011 ± 0.0006	0.8151 ± 0.0002	0.6849 ± 0.0005	0.8013 ± 0.0003
DKT+	0.7174 ± 0.0012	0.6947 ± 0.0008	0.8108 ± 0.0004	0.6834 ± 0.0010	0.8035 ± 0.0005
DKT-F	–	0.7069 ± 0.0005	0.8203 ± 0.0004	0.6897 ± 0.0009	0.8072 ± 0.0008
KQN	0.7216 ± 0.0007	0.7011 ± 0.0003	0.8145 ± 0.0004	0.6880 ± 0.0011	0.7993 ± 0.0006
SKVMN	0.7163 ± 0.0017	0.6871 ± 0.0003	0.8074 ± 0.0002	0.6829 ± 0.0008	0.7766 ± 0.0017
Deep-IRT	0.7189 ± 0.0009	0.7006 ± 0.0006	0.8149 ± 0.0002	0.6841 ± 0.0004	0.8026 ± 0.0009
qDKT	0.6711 ± 0.0030	0.7274 ± 0.0007	0.8284 ± 0.0003	0.6971 ± 0.0007	0.7258 ± 0.0005
GKT	0.7125 ± 0.0034	0.6842 ± 0.0192	0.8137 ± 0.0003	0.6953 ± 0.0010	0.7865 ± 0.0044
SAKT	0.7112 ± 0.0016	0.6803 ± 0.0010	0.8081 ± 0.0005	0.6643 ± 0.0012	0.7936 ± 0.0004
SAINT	0.6948 ± 0.0056	0.7129 ± 0.0014	0.8161 ± 0.0008	0.6606 ± 0.0106	0.7784 ± 0.0015
AKT	0.7289 ± 0.0009	0.7236 ± 0.0002	0.8241 ± 0.0002	0.7033 ± 0.0004	0.7999 ± 0.0016
RKT	0.7204 ± 0.0029	0.7232 ± 0.0002	0.8260 ± 0.0006	0.6949 ± 0.0015	0.7933 ± 0.0020
ATKT	0.7296 ± 0.0016	0.7029 ± 0.0003	0.8187 ± 0.0003	0.6926 ± 0.0003	0.8040 ± 0.0004
HawkesKT	0.7110 ± 0.0015	0.6405 ± 0.0006	0.7711 ± 0.0026	0.6471 ± 0.0009	0.7849 ± 0.0015
DIMKT	0.7286 ± 0.0018	0.7299 ± 0.0006	0.8287 ± 0.0004	0.7315 ± 0.0007	0.8112 ± 0.0015
AT-DKT	0.7214 ± 0.0008	0.7136 ± 0.0002	0.8193 ± 0.0004	0.7012 ± 0.0014	0.8100 ± 0.0013
simpleKT	0.7281 ± 0.0011	0.7319 ± 0.0003	0.8238 ± 0.0001	0.7112 ± 0.0014	0.8034 ± 0.0005
QIKT	0.7345 ± 0.0007	0.7318 ± 0.0006	0.8286 ± 0.0004	0.7224 ± 0.0010	**0.8220 ± 0.0015**
DTransformer	0.7260 ± 0.0010	0.7278 ± 0.0004	0.8230 ± 0.0004	0.7046 ± 0.0008	0.7959 ± 0.0022
FoLiBiKT	0.7343 ± 0.0024	0.7319 ± 0.0006	0.8231 ± 0.0005	0.7212 ± 0.0005	0.8031 ± 0.0007
extraKT	0.7294 ± 0.0022	0.7325 ± 0.0004	0.8255 ± 0.0004	0.7099 ± 0.0020	0.7985 ± 0.0019
Mamba4KT	0.6557 ± 0.0012	0.7023 ± 0.0002	0.8199 ± 0.0002	0.7040 ± 0.0004	0.7229 ± 0.0004
MCKT	0.6683 ± 0.0023	0.7197 ± 0.0004	0.8247 ± 0.0004	0.7078 ± 0.0006	0.7239 ± 0.0007
MTKT	–	0.7300 ± 0.0003	0.8263 ± 0.0001	0.7116 ± 0.0013	0.8098 ± 0.0017
LefoKT	0.7307 ± 0.0006	0.7299 ± 0.0004	0.8259 ± 0.0001	0.7046 ± 0.0010	0.8044 ± 0.0010
FlucKT	0.7318 ± 0.0012	0.7318 ± 0.0007	0.8248 ± 0.0006	0.7145 ± 0.0022	0.8118 ± 0.0016
UKT	0.7360 ± 0.0014	0.7228 ± 0.0013	0.8209 ± 0.0004	0.7143 ± 0.0024	0.8093 ± 0.0037
DenoiseKT	0.7363 ± 0.0024	0.7287 ± 0.0004	0.8295 ± 0.0001	0.7193 ± 0.0017	–
FA-KT	–	**0.7334 ± 0.0005**	0.8277 ± 0.0004	0.7226 ± 0.0010	0.8093 ± 0.0014
DualKT (ours)	**0.7390 ± 0.0008**	0.7333 ± 0.0004	**0.8319 ± 0.0003**	**0.7360 ± 0.0015**	0.8177 ± 0.0017

C.5 Unified-TPE Details

Figure 8 reports ACC under the same global TPE budget as the AUC comparison in Figure 3.

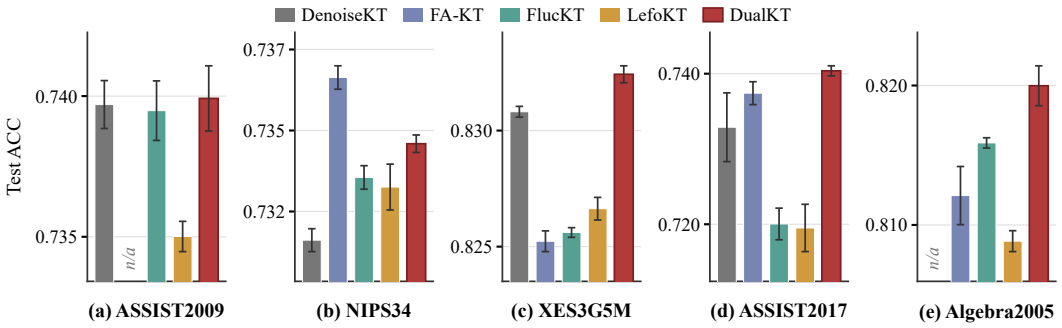


Figure 8: Test ACC under unified global TPE optimization. Bars and error bars show five-fold means ± one standard deviation.

C.6 Long-Sequence ACC Results

Figure 9 complements the main-text AUC analysis with classification accuracy over the same sequence-length sweep.

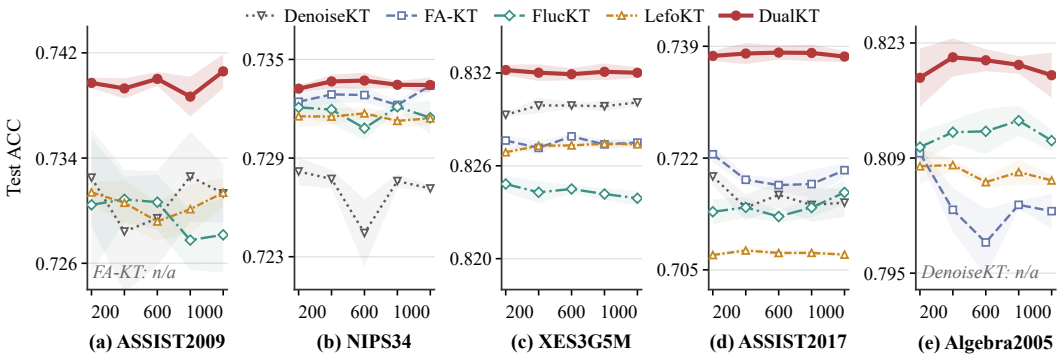


Figure 9: Long-sequence ACC robustness across five datasets. Curves and bands show five-fold mean test ACC $\pm$ one standard deviation as the maximum sequence length increases from 200 to 1,000.