BRIDGEFLOWKT: CONTINUOUS-TIME LATENT BELIEF-STATE DYNAMICS WITH QUERY-CONDITIONED UNCERTAINTY FOR KNOWLEDGE TRACING

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge tracing is commonly formulated as next-response prediction over ordered student interaction logs. Existing time-aware models often use elapsed time as an additional feature, gate, attention bias, or forgetting signal. This paper studies a complementary question: what changes if elapsed time acts directly on the learner state before the next response is predicted? We propose BRIDGEFLOWKT, a continuous-time knowledge tracing model that represents each learner by latent mean and uncertainty coordinates. Between observed attempts, both coordinates are transported by an exact exponential dynamics; at an observed response, a bounded event bridge updates the state; before prediction, the target question reads the pre-event state through query-conditioned mean and uncertainty interactions. This design keeps the roles of time, response evidence, and question-specific uncertainty explicit while still allowing a closed-form event scan. We analyze the resulting system through semigroup transport, stability under event perturbations, separation of histories with identical event order but different time gaps, uncertainty-channel separability, and nested mechanism ablations. Experiments on five benchmark datasets show that BRIDGEFLOWKT is competitive under matched per-dataset comparisons, improves consistently over the matched uncertainty-aware baseline, obtains its clearest aggregate temporal gain on ASSIST2012 with dataset-dependent evidence elsewhere, and benefits reliably from query-conditioned uncertainty interactions in matched ablations.

1 INTRODUCTION

Knowledge tracing (KT) estimates a student's evolving mastery from a sequence of exercise attempts and predicts future correctness (Corbett & Anderson, 1994; Piech et al., 2015). Standard KT objectives are usually written over ordered interaction logs, and many time-aware variants enrich this order with elapsed-time features, forgetting signals, temporal gates, or point-process effects. These methods show that timestamps are useful side information. Our focus is different: we ask whether the learner state itself can be constructed as a continuous-time belief path whose pre-response state is obtained by an explicit transport operator before the next event is assimilated.

This framing leads to a system-level design problem rather than a single feature-engineering choice. Between attempts, the state should evolve in physical time according to a specified flow; at attempts, response evidence should enter as a bounded event update; at prediction time, the readout should query the pre-event belief rather than only a discrete hidden vector. Uncertainty is important in this system because mastery is latent and responses are noisy. Classical BKT already treated uncertainty as central (Corbett & Anderson, 1994), and UKT develops uncertainty-aware neural representations and matching (Cheng et al., 2025). Our contribution is to make uncertainty a transported coordinate of the continuous-time belief state and to let the queried item determine how that coordinate affects prediction.

This motivates a belief-process view of KT: the learner state is a path in continuous time, responses are impulses that perturb the path, and prediction is a query-specific functional of the pre-event belief. The intended contribution is the construction of a coherent continuous-time modeling and theory

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.

stack for KT, with query-conditioned uncertainty serving as an important analyzable channel inside that stack.

Contributions. We make four contributions.

- We formulate KT as a continuous-time belief-flow system in which the learner state is explicitly transported between observed events.
- We propose BRIDGEFLOWKT, combining exact exponential transport, bounded event bridge updates, and a query-conditioned readout over mean and uncertainty coordinates.
- We prove theoretical properties that map directly to the implementation: exact semigroup transport, contraction, time-gap separability, uncertainty-channel separability, nested ablation hypothesis classes, and consistency of auxiliary losses.
- We design an empirical program centered on matched predictive comparison, time-gap ablations, uncertainty-channel ablations, auxiliary-loss tuning, and readout structure diagnostics.

2 RELATED WORK

Classical and neural KT. KT began with latent probabilistic mastery models such as BKT (Corbett & Anderson, 1994) and was later scaled with recurrent, memory, and attention-based neural architectures, including DKT, DKVMN, SAKT, AKT, and SimpleKT (Piech et al., 2015; Zhang et al., 2017; Pandey & Karypis, 2019; Ghosh et al., 2020; Liu et al., 2023b). These models are strong predictors, but their state evolution is usually indexed by interaction order. BRIDGEFLOWKT keeps the sequential prediction objective but makes the learner state a continuous-time belief path.

Temporal and process-aware KT. Several KT models explicitly encode forgetting, elapsed time, learning behavior, or process consistency (Nagatani et al., 2019; Wang et al., 2021; Shen et al., 2021; Xu et al., 2023; Huang et al., 2026). These works motivate our focus on physical time and show that timestamps can be useful. BRIDGEFLOWKT differs in where time enters: instead of only conditioning a discrete update or readout, elapsed time transports the whole mean–uncertainty state by an exact exponential semigroup before the next event is read.

Question, concept, and interpretable structure. Question-centric, graph-based, difficulty-aware, and interpretable KT models show that item structure and diagnostic readouts matter for practical KT (Chen et al., 2023; Sonkar et al., 2020; Yang et al., 2020; Nakagawa et al., 2019; Cui et al., 2024b; Shen et al., 2022; Sun et al., 2024). Our readout follows this lesson by making prediction query-conditioned, but the distinctive object is the transported belief chart rather than a new question representation alone.

Uncertainty-aware KT. UKT introduces stochastic representations and uncertainty-aware matching for KT (Cheng et al., 2025). Our contribution is to make uncertainty a transported coordinate of the continuous-time belief state and to let the queried item determine how that coordinate affects prediction.

3 PROBLEM FORMULATION

3.1 OBSERVED HISTORIES AND FILTRATIONS

For a student u, let

$$\mathcal{H}_u^n = \{(q_{u,i}, c_{u,i}, y_{u,i}, t_{u,i})\}_{i=1}^n, \tag{1}$$

where $q_{u,i} \in \mathcal{Q}$ is a question, $c_{u,i}$ denotes its associated concept set or support vector, $y_{u,i} \in \{0,1\}$ is correctness, and $t_{u,i} \in \mathbb{R}_+$ is the timestamp. The learner's information before observing $y_{u,i}$ is represented by the filtration

$$\mathcal{F}_{u,i}^- = \sigma\big((q_{u,j}, c_{u,j}, y_{u,j}, t_{u,j})_{j<i}, q_{u,i}, c_{u,i}, t_{u,i}\big), \tag{2}$$

and after observing the response by

$$\mathcal{F}_{u,i}^+ = \sigma\big(\mathcal{F}_{u,i}^-, y_{u,i}\big). \tag{3}$$

Here $\sigma(\cdot)$ denotes the generated sigma-algebra; the uncertainty scale below is written in bold as $\boldsymbol{\sigma}(\cdot)$. The elapsed time is measured in hours. We set $t_{u,0} = t_{u,1}$ so that the first gap is zero, and use a fixed maximum $\tau_{\max}$ for elapsed-time clipping:

$$\Delta t_{u,i}^{\text{raw}} = \frac{\max(t_{u,i} - t_{u,i-1}, 0)}{3600}, \qquad \Delta t_{u,i} = \min\{\Delta t_{u,i}^{\text{raw}}, \tau_{\max}\}. \tag{4}$$

Each question has a sparse concept-support vector $\mathbf{s}(q) \in \{0,1\}^K$ with normalized support

$$\bar{\mathbf{s}}(q) = \frac{\mathbf{s}(q)}{\max(\|\mathbf{s}(q)\|_1, 1)}. \tag{5}$$

The prediction target is

$$p_{u,i}^{\star} = \mathbb{P}(y_{u,i} = 1 \mid \mathcal{F}_{u,i}^{-}). \tag{6}$$

A KT model outputs $\hat{p}_{u,i} = g_\theta(\mathcal{F}_{u,i}^{-})$ and is evaluated by proper scoring rules and ranking metrics.

3.2 Belief-state modeling objective

A belief-state KT model maintains a state $B_{u,i}^{+}$ that is $\mathcal{F}_{u,i}^{+}$-measurable after event i. Writing its chart coordinates as $\mathbf{z}_{u,i}^{\pm}$, before event $i+1$ it deterministically transports the state using elapsed time and then predicts:

$$\mathbf{z}_{u,i+1}^{-} = \Phi_{\Delta t_{u,i+1}}(\mathbf{z}_{u,i}^{+}), \tag{7}$$

$$\hat{p}_{u,i+1} = \text{sigmoid}\big(\Gamma_\theta(\mathbf{z}_{u,i+1}^{-}, q_{u,i+1}, \Delta t_{u,i+1})\big). \tag{8}$$

The scientific question is whether the map $\Phi_{\Delta t}$ should be a genuine continuous-time transport operator rather than an implicit transition hidden inside a discrete sequence encoder.

4 BridgeFlowKT: A Belief-Flow Calculus

The model is easiest to understand as a calculus over latent belief trajectories. A student history induces a càdlàg path: between observed attempts the belief follows a continuous-time contraction flow, and at an attempt the belief receives a bounded jump determined by the event. Prediction is not a generic classifier on the state; it is a query-specific chart that exposes which parts of the mean and uncertainty channels are used.

4.1 State chart and uncertainty geometry

Let

$$\mathbf{z}_i^{\pm} = \begin{bmatrix} \boldsymbol{\mu}_i^{\pm} \\ \boldsymbol{\rho}_i^{\pm} \end{bmatrix} \in \mathbb{R}^{2d}, \qquad \boldsymbol{\sigma}(\boldsymbol{\rho}) = \text{softplus}(\boldsymbol{\rho}) + \sigma_{\min}\mathbf{1}, \qquad \Sigma(\boldsymbol{\rho}) = \text{diag}(\boldsymbol{\sigma}(\boldsymbol{\rho})^2). \tag{9}$$

These $\mathbf{z}_i^{\pm}$ are the chart coordinates used for the corresponding belief states throughout the model. Here $\boldsymbol{\mu}$ is the mean-like mastery coordinate, while $\boldsymbol{\rho}$ is an unconstrained uncertainty coordinate whose positive scale is $\boldsymbol{\sigma}(\boldsymbol{\rho})$, with fixed floor $\sigma_{\min} > 0$. The diagonal matrix $\Sigma(\boldsymbol{\rho})$ is not used to claim a fully Bayesian posterior; it is a structured uncertainty chart that can be transported, queried, and ablated. The clipped reachable set is

$$\mathcal{Z}_c = [-c_\mu, c_\mu]^d \times [-c_\rho, c_\rho]^d, \qquad \Pi_c(\mathbf{z}) = \text{clip}(\mathbf{z}, \mathcal{Z}_c), \tag{10}$$

where $c_\mu, c_\rho > 0$ are clipping radii and clipping is coordinatewise.

4.2 Continuous-time semigroup with event jumps

Define the equilibrium and block-diagonal decay operator

$$\mathbf{z}_{\text{eq}} = \begin{bmatrix} \boldsymbol{\mu}_{\text{eq}} \\ \boldsymbol{\rho}_{\text{eq}} \end{bmatrix}, \qquad \Lambda = \text{blkdiag}(\Lambda_\mu, \Lambda_\rho), \qquad \Lambda_\mu = \text{diag}(\boldsymbol{\lambda}_\mu), \quad \Lambda_\rho = \text{diag}(\boldsymbol{\lambda}_\rho), \tag{11}$$

with learnable unconstrained parameters $\mathbf{a}_\mu, \mathbf{a}_\rho$ and a small floor $\epsilon > 0$, so that $\boldsymbol{\lambda}_\mu = \mathrm{softplus}(\mathbf{a}_\mu) + \epsilon\mathbf{1}$ and $\boldsymbol{\lambda}_\rho = \mathrm{softplus}(\mathbf{a}_\rho) + \epsilon\mathbf{1}$. All continuous-time arguments τ and t below use the same hour scale as the clipped gaps Δt_i. Between events,

$$\frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} = -\Lambda(\mathbf{z}(t) - \mathbf{z}_{\mathrm{eq}}), \qquad S_\tau = \exp(-\tau\Lambda), \tag{12}$$

and the exact transport map is

$$\Phi_\tau(\mathbf{z}) = \mathbf{z}_{\mathrm{eq}} + S_\tau(\mathbf{z} - \mathbf{z}_{\mathrm{eq}}) = \begin{bmatrix} \boldsymbol{\mu}_{\mathrm{eq}} + e^{-\tau\Lambda_\mu}(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{eq}}) \\ \boldsymbol{\rho}_{\mathrm{eq}} + e^{-\tau\Lambda_\rho}(\boldsymbol{\rho} - \boldsymbol{\rho}_{\mathrm{eq}}) \end{bmatrix}. \tag{13}$$

For an observed event $(q_i, y_i, \Delta t_i)$, the event feature is

$$\mathbf{e}_i = \big[\mathbf{e}_q(q_i); \mathbf{e}_c(q_i); \mathbf{e}_y(y_i); y_i; \log(1 + \Delta t_i)\big], \tag{14}$$

where $\mathbf{e}_q$, $\mathbf{e}_c$, and $\mathbf{e}_y$ denote learned question, concept-side, and response embeddings, respectively. The event encoder M_{evt} is a learned map from $\mathbf{e}_i$ to the bridge hidden state. The bridge displacement, controlled by positive jump scales δ_μ and δ_ρ, is

$$\begin{bmatrix} \mathbf{h}_{\mu,i} \\ \mathbf{h}_{\rho,i} \end{bmatrix} = M_{\mathrm{evt}}(\mathbf{e}_i), \qquad \Delta\mathbf{z}_i = D_\theta(\mathbf{e}_i) = \begin{bmatrix} \delta_\mu \tanh(\mathbf{h}_{\mu,i}) \\ \delta_\rho \tanh(\mathbf{h}_{\rho,i}) \end{bmatrix}. \tag{15}$$

The belief recurrence is therefore the projected jump-flow

$$\mathbf{z}_i^- = \Phi_{\Delta t_i}(\mathbf{z}_{i-1}^+), \qquad \mathbf{z}_i^+ = \Pi_c(\mathbf{z}_i^- + \Delta\mathbf{z}_i). \tag{16}$$

This compact form separates three roles: S_τ carries physical-time memory, D_θ assimilates response evidence, and Π_c keeps the chart in a stable compact region.

Equivalently, before projection, the centered path $\tilde{\mathbf{z}}(t) = \mathbf{z}(t) - \mathbf{z}_{\mathrm{eq}}$ is the solution of the impulsive system

$$\mathrm{d}\tilde{\mathbf{z}}(t) = -\Lambda\tilde{\mathbf{z}}(t)\mathrm{d}t + \sum_i \Delta\mathbf{z}_i\, \delta_{t_i}(\mathrm{d}t), \qquad \tilde{\mathbf{z}}(t) = S_{t-t_0}\tilde{\mathbf{z}}(t_0^+) + \sum_{i:t_i \le t} S_{t-t_i}\Delta\mathbf{z}_i. \tag{17}$$

Here $\delta_{t_i}(\mathrm{d}t)$ is a Dirac measure at event time t_i in the hour-rescaled clock. This variation-of-constants form is the continuous-time object represented by the finite recurrence.

4.3 Exact scan identity

Ignoring the projection only for algebraic exposition, define centered states $\tilde{\mathbf{z}}_i^\pm = \mathbf{z}_i^\pm - \mathbf{z}_{\mathrm{eq}}$, cumulative time $T_i = \sum_{r=1}^i \Delta t_r$, and $P_i = \exp(-T_i\Lambda)$. Then

$$\tilde{\mathbf{z}}_i^- = P_i\left(\tilde{\mathbf{z}}_0^+ + \sum_{j<i} P_j^{-1}\Delta\mathbf{z}_j\right), \qquad \tilde{\mathbf{z}}_i^+ = P_i\left(\tilde{\mathbf{z}}_0^+ + \sum_{j\le i} P_j^{-1}\Delta\mathbf{z}_j\right). \tag{18}$$

Thus the continuous-time model admits an exact vectorized scan: exponentials encode elapsed time, and cumulative sums encode event evidence. The implemented chunked scan is a numerically stabilized version of Equation (18), with clipping inserted after jumps.

4.4 Query-conditioned uncertainty chart

For a target question q, define the query covector

$$\mathbf{r}_q = \tanh\left(W_r[\mathbf{e}_q(q); \mathbf{e}_c(q)]\right) \in \mathbb{R}^d. \tag{19}$$

Here W_r, $W_{s,\mu}$, $W_{s,\rho}$, W_c, and b_q denote learned readout parameters of compatible dimensions; W_μ, W_σ, $\mathbf{b}$, and η are optional precision-tempering parameters used below. For $\mathbf{z}^- = (\boldsymbol{\mu}^-, \boldsymbol{\rho}^-)$ and $\boldsymbol{\sigma}^- = \boldsymbol{\sigma}(\boldsymbol{\rho}^-)$, introduce two query operators

$$A_q^\mu = \begin{bmatrix} \mathbf{r}_q^\top/\sqrt{d} \\ \mathbf{1}^\top/d \\ \mathrm{diag}(\mathbf{r}_q) \end{bmatrix}, \qquad A_q^\sigma = \begin{bmatrix} |\mathbf{r}_q|^\top/\sqrt{d} \\ \mathbf{1}^\top/d \\ \mathrm{diag}(|\mathbf{r}_q|) \end{bmatrix}, \tag{20}$$

$$M_q(\boldsymbol{\mu}^-) = A_q^\mu\boldsymbol{\mu}^-, \qquad U_q(\boldsymbol{\rho}^-) = A_q^\sigma\boldsymbol{\sigma}(\boldsymbol{\rho}^-). \tag{21}$$

The state-summary coordinates are block separated as

$$H_s(\mathbf{z}^-) = \big[\tanh(W_{s,\mu}\boldsymbol{\mu}^-); \tanh(W_{s,\rho}\boldsymbol{\rho}^-)\big], \tag{22}$$

so uncertainty ablations can remove the $\boldsymbol{\rho}$-summary block without having to separate mixed coordinates. The full readout feature map is

$$\psi_q(\mathbf{z}^-, \tau) = \big[M_q(\boldsymbol{\mu}^-); U_q(\boldsymbol{\rho}^-); \log(1+\tau); H_s(\mathbf{z}^-); \mathbf{e}_q(q); \mathbf{e}_c(q); \mathbf{r}_q\big]. \tag{23}$$

A concept-side scaffold and residual neural chart give the logit

$$\Gamma_\theta(\mathbf{z}^-, q, \tau) = \big\langle W_c\boldsymbol{\mu}^-, \bar{\mathbf{s}}(q)\big\rangle + \gamma f_\theta(\psi_q(\mathbf{z}^-, \tau)) + b_q, \qquad \hat{p}(\mathbf{z}^-, q, \tau) = \mathrm{sigmoid}\big(\Gamma_\theta(\mathbf{z}^-, q, \tau)\big). \tag{24}$$

Here $\gamma > 0$ is the residual scale and b_q is a learned question bias. The uncertainty contribution is not a single confidence scalar. Its local differential is structured by

$$\mathrm{d}U_q = A_q^\sigma \mathrm{diag}(\mathrm{softplus}'(\boldsymbol{\rho}^-))\,\mathrm{d}\boldsymbol{\rho}^-, \tag{25}$$

which shows explicitly how a change in uncertainty is filtered through the query. Precision-tempered heads use the same chart but normalize mean interactions by learned uncertainty precision:

$$\tilde{\mathbf{u}}_q = \big(W_\mu(\boldsymbol{\mu}^- \odot \mathbf{r}_q) + \mathbf{b}\big) \odot \big(\mathbf{1} + \mathrm{softplus}\,\big(W_\sigma(\boldsymbol{\sigma}^- \odot |\mathbf{r}_q|) + \eta \log(1+\tau)\big)\big)^{-1/2}. \tag{26}$$

4.5 Mask calculus for uncertainty ablations

Let M be a diagonal mask acting on coordinates of ψ_q. Because the state summary in Equation (22) is block separated, each uncertainty ablation is a precise projection of the same feature chart:

$$\Gamma_{\theta,M}(\mathbf{z}, q, \tau) = \langle W_c\boldsymbol{\mu}, \bar{\mathbf{s}}(q)\rangle + \gamma f_\theta(M\psi_q(\mathbf{z}, \tau)) + b_q. \tag{27}$$

The main masks are

$$\begin{aligned} M_{\text{full}}\psi_q &= \psi_q, \\ M_{\mu\text{-only}}\psi_q &= \psi_q \text{ with } U_q(\boldsymbol{\rho}) \text{ and the } \boldsymbol{\rho}\text{-summary block of } H_s(\mathbf{z}) \text{ zeroed}, \\ M_{\text{no scalar}}\psi_q &= \psi_q \text{ with } \langle \boldsymbol{\sigma}(\boldsymbol{\rho}), |\mathbf{r}_q|\rangle/\sqrt{d} \text{ and } \mathbf{1}^\top\boldsymbol{\sigma}(\boldsymbol{\rho})/d \text{ zeroed}, \\ M_{\text{no interaction}}\psi_q &= \psi_q \text{ with } \boldsymbol{\sigma}(\boldsymbol{\rho}) \odot |\mathbf{r}_q| \text{ and its descendants zeroed}. \end{aligned} \tag{28}$$

Because all masks operate on one chart, ablation comparisons test information pathways rather than unrelated architectures.

4.6 Training objective as scoring plus path regularization

For an anchor i and target $j > i$, let $\tau_{i\to j}^{\text{raw}} = \max(t_j - t_i, 0)/3600$ and $\tau_{i\to j} = \min\{\tau_{i\to j}^{\text{raw}}, \tau_{\max}\}$ denote the clipped elapsed horizon between them in the same units as Δt, and define the transported prediction

$$p_{i\to j} = \mathrm{sigmoid}\left(\Gamma_\theta\left(\Phi_{\tau_{i\to j}}(\mathbf{z}_i^+), q_j, \tau_{i\to j}\right)\right). \tag{29}$$

Let $\mathcal{I}_1$ be next-step pairs, $\mathcal{A}$ be selected anytime anchor-target pairs, $\mathcal{C}$ be consistency triples (i,k,j) with $i < k < j$, and $\mathcal{J}$ be the event indices included in the bridge-energy penalty. With binary cross entropy abbreviated as BCE, $\Delta\mathbf{z} = (\Delta\boldsymbol{\mu}, \Delta\boldsymbol{\rho})$, and uncertainty weight $\alpha_\rho > 0$, define the metric

$$\|\Delta\mathbf{z}\|_G^2 = \Delta\mathbf{z}^\top G\Delta\mathbf{z} = \|\Delta\boldsymbol{\mu}\|_2^2 + \alpha_\rho\|\Delta\boldsymbol{\rho}\|_2^2, \qquad G = \mathrm{blkdiag}(I_d, \alpha_\rho I_d), \qquad w_i = 1 + \log(1+\Delta t_i), \tag{30}$$

and the loss terms are

$$\begin{aligned} \mathcal{L}_{\text{bce}} &= \frac{1}{|\mathcal{I}_1|}\sum_{(i,j)\in\mathcal{I}_1} \mathrm{BCE}(y_j, p_{i\to j}), & \mathcal{L}_{\text{any}} &= \frac{1}{|\mathcal{A}|}\sum_{(i,j)\in\mathcal{A}} \mathrm{BCE}(y_j, p_{i\to j}), \\ \mathcal{L}_{\text{cons}} &= \frac{1}{|\mathcal{C}|}\sum_{(i,k,j)\in\mathcal{C}} (p_{i\to j} - p_{k\to j})^2, & \mathcal{L}_{\text{energy}} &= \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} \frac{\|\Delta\mathbf{z}_i\|_G^2}{w_i}, \\ \mathcal{L}_{\text{flow}} &= \mathbb{E}_{i,\alpha}\left\|v_{\text{flow}}\left((1-\alpha)\mathbf{z}_i^- + \alpha\mathbf{z}_i^+, \mathbf{e}_i, \alpha\right) - \Delta\mathbf{z}_i\right\|_G^2, & \alpha &\sim \mathrm{Unif}(0,1). \end{aligned} \tag{31}$$

Here v_{flow} denotes the learned vector field used by the flow-matching loss. The optimized objective is

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{bce}} &+ \lambda_{\text{any}}\mathcal{L}_{\text{any}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} \\ &+ \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}}. \end{aligned} \tag{32}$$

5 THEORETICAL PROPERTIES OF THE BELIEF-FLOW CALCULUS

The purpose of the theory is not to decorate the architecture. Each result isolates one structural claim that is later tested by an ablation or probe.

5.1 FLOW ALGEBRA AND EXACT COMPUTATION

Theorem 1 (Exponential semigroup). *The maps $\{\Phi_\tau\}_{\tau\geq 0}$ defined in Equation* (13) *are the unique solution operators of Equation* (12)*. They satisfy*

$$\Phi_0 = \mathrm{Id}, \qquad \Phi_{\tau+s} = \Phi_\tau \circ \Phi_s, \qquad \|\Phi_\tau(\mathbf{z}) - \Phi_\tau(\mathbf{z}')\|_2 \leq e^{-\lambda_{\min}\tau}\|\mathbf{z}-\mathbf{z}'\|_2, \tag{33}$$

where $\lambda_{\min} = \min_m\{\lambda_{\mu,m}, \lambda_{\rho,m}\}$.

Theorem 2 (Closed-form scan). *Without projection, the jump-flow recurrence in Equation* (16) *is exactly equivalent to the scan identity in Equation* (18)*. Hence a length-n belief path can be computed from $(P_i)_{i=1}^n$ and cumulative sums of $P_i^{-1}\Delta\mathbf{z}_i$, without numerical ODE solving.*

5.2 PATHWISE STABILITY

Proposition 1 (Perturbation contraction). *Consider two paths $\{\mathbf{z}_i^\pm\}$ and $\{\bar{\mathbf{z}}_i^\pm\}$ with the same gaps and displacements $\Delta\mathbf{z}_i, \Delta\bar{\mathbf{z}}_i$, and assume the projection Π_c is Euclidean nonexpansive. Then*

$$\|\mathbf{z}_n^+ - \bar{\mathbf{z}}_n^+\|_2 \leq e^{-\lambda_{\min}T_n}\|\mathbf{z}_0^+ - \bar{\mathbf{z}}_0^+\|_2 + \sum_{j=1}^{n} e^{-\lambda_{\min}(T_n - T_j)}\|\Delta\mathbf{z}_j - \Delta\bar{\mathbf{z}}_j\|_2. \tag{34}$$

Thus early event perturbations are exponentially discounted according to physical elapsed time, not merely according to sequence distance.

Proposition 2 (Compact reachable set). *With clipping in Equation* (10)*, every post-event state lies in $\mathcal{Z}_c$. Every pre-event state lies in the coordinatewise compact hull of $\mathcal{Z}_c \cup \{\mathbf{z}_{\mathrm{eq}}\}$, namely the axis-aligned compact box obtained by adjoining $\mathbf{z}_{\mathrm{eq}}$ to the coordinate ranges of $\mathcal{Z}_c$. Consequently Γ_θ is bounded on reachable states whenever its neural components have finite weights.*

5.3 TEMPORAL RESPONSE AND GAP ALIASING

Theorem 3 (Temporal logit derivative). *Fix a post-event state $\mathbf{z}^+$ and query q. Let $g_q(\tau) = \Gamma_\theta(\Phi_\tau(\mathbf{z}^+), q, \tau)$, where the explicit scalar gap feature is $r(\tau) = \log(1+\tau)$. In the derivative below, ∂_r means differentiation with respect to this log-gap coordinate. Wherever Γ_θ is differentiable,*

$$\frac{\mathrm{d}g_q(\tau)}{\mathrm{d}\tau} = -\nabla_{\mathbf{z}}\Gamma_\theta(\Phi_\tau(\mathbf{z}^+), q, \tau)^\top \Lambda S_\tau(\mathbf{z}^+ - \mathbf{z}_{\mathrm{eq}}) + \frac{1}{1+\tau}\,\partial_r\Gamma_\theta(\Phi_\tau(\mathbf{z}^+), q, \tau). \tag{35}$$

Consequently, for any $\tau_1 < \tau_2$,

$$|g_q(\tau_2) - g_q(\tau_1)| \leq \int_{\tau_1}^{\tau_2} \left(\|\nabla_{\mathbf{z}}\Gamma_\theta(\Phi_s(\mathbf{z}^+), q, s)\|_2 \, \|\Lambda S_s(\mathbf{z}^+ - \mathbf{z}_{\mathrm{eq}})\|_2 + \frac{|\partial_r\Gamma_\theta(\Phi_s(\mathbf{z}^+), q, s)|}{1+s} \right) \mathrm{d}s. \tag{36}$$

Theorem 4 (Gap-aliasing separation). *Any gap-agnostic predictor assigns identical outputs to histories with the same ordered question-response sequence and non-temporal shortcuts. In contrast, if $\mathbf{z}^+ \neq \mathbf{z}_{\mathrm{eq}}$ and for some query q the derivative in Equation* (35) *is not identically zero on an interval, then* BRIDGEFLOWKT *separates a positive-measure set of gap pairs in that interval.*

5.4 UNCERTAINTY-CHANNEL SEPARABILITY

Definition 1 (Mean quotient). Two states $\mathbf{z} = (\boldsymbol{\mu}, \boldsymbol{\rho})$ and $\mathbf{z}' = (\boldsymbol{\mu}', \boldsymbol{\rho}')$ are mean-equivalent if $\boldsymbol{\mu} = \boldsymbol{\mu}'$. A mean-only readout is any readout that factors through the quotient map $\pi_\mu(\mathbf{z}) = \boldsymbol{\mu}$.

Theorem 5 (Query-conditioned uncertainty separation). *Let $\mathbf{z} = (\boldsymbol{\mu}, \boldsymbol{\rho})$ and $\mathbf{z}' = (\boldsymbol{\mu}, \boldsymbol{\rho}')$ be mean-equivalent. Mean-only readouts cannot distinguish them. The full chart distinguishes them for query q whenever*

$$A_q^\sigma\left(\boldsymbol{\sigma}(\boldsymbol{\rho}) - \boldsymbol{\sigma}(\boldsymbol{\rho}')\right) \neq \mathbf{0}, \qquad A_q^\sigma = \begin{bmatrix} |\mathbf{r}_q|^\top/\sqrt{d} \\ \mathbf{1}^\top/d \\ \mathrm{diag}(|\mathbf{r}_q|) \end{bmatrix}, \tag{37}$$

and f_θ has nonzero local derivative in the resulting feature direction. Locally, the uncertainty sensitivity is

$$\nabla_{\boldsymbol{\rho}}\Gamma_\theta = \gamma\,(D_U f_\theta)\,A_q^\sigma \mathrm{diag}(\mathrm{softplus}'(\boldsymbol{\rho})) + \gamma\,(D_h f_\theta)\,D_{\boldsymbol{\rho}} H_s(\mathbf{z}), \tag{38}$$

where $D_U f_\theta$ and $D_h f_\theta$ denote the readout derivatives with respect to uncertainty-chart and state-summary coordinates.

Proposition 3 (Nested ablation classes)**.** *For each mask M in Equation* (28)*, let $\mathcal{M}_M = \{\Gamma_{\theta,M} : \theta\}$ and let $\mathcal{M}_{\mathrm{full}}$ use $M = I$. If the first layer of f_θ can set arbitrary incoming weights to zero, then*

$$\mathcal{M}_M \subseteq \mathcal{M}_{\mathrm{full}}, \qquad \inf_{\Gamma\in\mathcal{M}_{\mathrm{full}}} \widehat{R}(\Gamma) \le \inf_{\Gamma\in\mathcal{M}_M} \widehat{R}(\Gamma) \tag{39}$$

for any empirical risk $\widehat{R}$, up to optimization and regularization effects.

5.5 Auxiliary losses as variational constraints

Proposition 4 (Properness of transported BCE)**.** *For any fixed anchor information $\mathcal{F}_i^+$ and target query-gap pair (q_j, τ), let $Y_j \in \{0, 1\}$ denote the random correctness at the target. The population risk*

$$R(p) = \mathbb{E}[-Y_j \log p - (1 - Y_j)\log(1 - p) \mid \mathcal{F}_i^+, q_j, \tau] \tag{40}$$

is uniquely minimized at

$$p^\star = \mathbb{P}(Y_j = 1 \mid \mathcal{F}_i^+, q_j, \tau). \tag{41}$$

Thus the anytime term adds proper transported prediction constraints rather than changing the population target.

Proposition 5 (Flow-matching projection)**.** *For $X_i = ((1-\alpha)\mathbf{z}_i^- + \alpha\mathbf{z}_i^+, \mathbf{e}_i, \alpha)$ and target vector $\boldsymbol{\xi}_i = \Delta\mathbf{z}_i$, the minimizer of $\mathbb{E}\|v(X_i) - \boldsymbol{\xi}_i\|_G^2$ over square-integrable vector fields is*

$$v^\star(X_i) = \mathbb{E}[\boldsymbol{\xi}_i \mid X_i]. \tag{42}$$

In the deterministic bridge used by BridgeFlowKT*, $\boldsymbol{\xi}_i$ is determined by $\mathbf{e}_i$, so the optimum equals the event displacement.*

Proposition 6 (Weighted energy controls path variation)**.** *For any finite event-index set H,*

$$\sum_{i\in H} \|\Delta\mathbf{z}_i\|_G \le \left(\sum_{i\in H} w_i\right)^{1/2} \left(\sum_{i\in H} \frac{\|\Delta\mathbf{z}_i\|_G^2}{w_i}\right)^{1/2}. \tag{43}$$

Thus the energy penalty is a weighted total-variation control on the jump component of the belief path.

6 Experiments

Our evaluation addresses four questions. First, is BridgeFlowKT competitive with strong KT baselines, especially UKT, under the AUC reporting used in this benchmark setting? Second, do direct ablations show that physical time and transported uncertainty matter? Third, do timestamp-quality and mechanism diagnostics clarify when aggregate AUC should be interpreted cautiously? Fourth, do optimization auxiliaries and readout structure help for the reasons claimed rather than through unrelated shortcuts?

6.1 Experimental protocol

All methods are trained under the same benchmark split protocol and the same basic optimization schedule unless an ablation explicitly changes one component. The main board uses ASSIST2012, ASSIST2017, Junyi2015, MOOCRadar-C746997, and XES3G5M (Feng et al., 2009; Pardos et al., 2013; Yu et al., 2023; Liu et al., 2023c). Mechanism ablations focus on ASSIST2012, ASSIST2017, and XES3G5M. Because many benchmark baselines are reported under dataset-specific model selection, Table 1 presents BridgeFlowKT under matched within-dataset tuning across all five datasets; separate appendix tables summarize the clean-dataset tuning confirmations and the shared-reference configuration used for ablations and cross-dataset coverage checks. We report means and standard deviations across evaluation splits and use conservative significance language because some comparator coverage remains constrained by dataset-specific feature requirements. Full claim-to-test mapping, dataset statistics and source links, baseline compatibility notes, metric definitions, and additional diagnostic material are deferred to Appendix B.

Table 1: Main predictive comparison in AUC. Entries are mean ± standard deviation across evaluation splits, and em dashes denote unavailable entries. Best results are bold and second-best results are underlined. The BRIDGEFLOWKT row reports the within-dataset tuned results included in this study.

Method	ASSIST2012	ASSIST2017	Junyi2015	MOOCRadar	XES3G5M
DKT	0.6627 ± 0.0094	0.7182 ± 0.0125	0.6838 ± 0.0120	0.7581 ± 0.0072	0.7604 ± 0.0007
DKVMN	0.6801 ± 0.0117	0.6980 ± 0.0070	0.6978 ± 0.0116	0.7521 ± 0.0059	0.7791 ± 0.0083
ATKT	0.6724 ± 0.0129	0.7183 ± 0.0115	0.6917 ± 0.0117	0.7560 ± 0.0103	0.7733 ± 0.0025
AKT	0.7492 ± 0.0102	0.7567 ± 0.0089	0.7692 ± 0.0069	0.7667 ± 0.0067	0.8191 ± 0.0099
ATDKT	0.7381 ± 0.0110	0.7821 ± 0.0091	0.7614 ± 0.0072	0.7701 ± 0.0096	0.8195 ± 0.0074
SimpleKT	0.7448 ± 0.0098	0.7518 ± 0.0059	0.7715 ± 0.0068	0.7619 ± 0.0102	0.8217 ± 0.0071
SparseKT	0.7426 ± 0.0105	0.7311 ± 0.0029	0.7704 ± 0.0069	0.7726 ± 0.0055	0.8171 ± 0.0065
QIKT	0.7448 ± 0.0116	0.7704 ± 0.0075	0.7616 ± 0.0075	0.7770 ± 0.0046	0.8239 ± 0.0077
qDKT	0.7502 ± 0.0099	0.7805 ± 0.0088	0.7660 ± 0.0068	$\underline{0.7785 \pm 0.0065}$	$\mathbf{0.8256 \pm 0.0077}$
LBKT	$\underline{0.7565 \pm 0.0141}$	$\mathbf{0.8162 \pm 0.0110}$	$\mathbf{0.7930 \pm 0.0048}$	—	—
LPKT	0.7479 ± 0.0130	$\underline{0.7877 \pm 0.0114}$	0.7611 ± 0.0091	—	—
DKTForget	0.7461 ± 0.0028	0.7263 ± 0.0114	0.7142 ± 0.0093	0.7402 ± 0.0091	0.7851 ± 0.0041
UKT	0.7223 ± 0.0112	0.6238 ± 0.0022	0.7298 ± 0.0145	0.7410 ± 0.0005	0.7933 ± 0.0058
BridgeFlowKT	$\mathbf{0.7674 \pm 0.0072}$	0.7466 ± 0.0007	$\underline{0.7717 \pm 0.0065}$	$\mathbf{0.7791 \pm 0.0050}$	$\underline{0.8250 \pm 0.0029}$

Table 2: Direct uncertainty ablations in AUC on the diagnostic subset. Entries are mean ± standard deviation across reported evaluation splits. The full row uses the matched no-anytime setting because the uncertainty ablations set the anytime weight to zero.

Variant	Unc. scalar	Unc. interaction	ASSIST2012	ASSIST2017	XES3G5M
full, no-anytime	yes	yes	0.7490 ± 0.0051	0.7520 ± 0.0041	0.8146 ± 0.0081
mu-only readout	no	no	0.7383 ± 0.0049	0.7435 ± 0.0029	0.8052 ± 0.0088
no scalar uncertainty	no	yes	0.7412 ± 0.0050	0.7460 ± 0.0024	0.8075 ± 0.0079
no interaction uncertainty	yes	no	0.7396 ± 0.0054	0.7445 ± 0.0031	0.8060 ± 0.0089
isolated interaction head	yes	yes	0.7351 ± 0.0069	0.7421 ± 0.0011	0.8035 ± 0.0082
isolated mu-only head	no	no	0.7324 ± 0.0059	0.7405 ± 0.0039	0.8014 ± 0.0085
isolated no-scalar head	no	yes	0.7360 ± 0.0061	0.7411 ± 0.0010	0.8020 ± 0.0077
isolated no-interaction head	yes	no	0.7343 ± 0.0041	0.7402 ± 0.0029	0.8003 ± 0.0091

6.2 MAIN COMPARISON

Table 1 presents matched comparisons. On ASSIST2012, tuned BRIDGEFLOWKT is strongest at 0.7674 ± 0.0072. On ASSIST2017, BRIDGEFLOWKT reaches 0.7466 ± 0.0007, improving over UKT while trailing the strongest attention- and behavior-aware baselines. On Junyi2015, BRIDGEFLOWKT reaches 0.7717 ± 0.0065, placing it in a tight top cluster with AKT, SimpleKT, and SparseKT while trailing LBKT. On MOOCRadar, tuned BRIDGEFLOWKT reaches the highest reported mean AUC at 0.7791 ± 0.0050, with a small margin over qDKT. On XES3G5M, tuned BRIDGEFLOWKT reaches 0.8250 ± 0.0029, remaining within the top cluster and close to qDKT. The matched UKT row is reported on all five datasets, and tuned BRIDGEFLOWKT exceeds UKT on all five. LPKT requires elapsed-time and use-time annotations, whereas LBKT additionally requires hint and attempt-count annotations; MOOCRadar-C746997 and XES3G5M do not provide this full annotation set, so the corresponding entries are unavailable. These results support competitive predictive performance across matched settings together with explicit temporal transport, strong uncertainty ablations, and matched auxiliary-supervision diagnostics. The shared-reference configuration used for ablations and cross-dataset coverage checks is described in Appendix B.

6.3 UNCERTAINTY ABLATIONS

Table 2 is the central mechanism table. Relative to the mean-only readout, the matched full row is higher by about 0.0107 on ASSIST2012, 0.0085 on ASSIST2017, and 0.0094 on XES3G5M. Removing either scalar or interaction uncertainty reduces performance on all three datasets, with the larger drop consistently coming from removing the query-conditioned interaction path. The isolated-head rows are lower still, which shows that the gains are not explained by head capacity alone. Overall, the uncertainty channel contributes materially, and the strongest part of that contribution comes from the query-conditioned interaction pathway, with the scalar uncertainty path providing an additional complementary gain. Calibration metrics and factor associations are reported in Appendix B, Tables 12 and 13.

Table 3: Time-gap confirmation across all five datasets on the diagnostic reference runs. Entries are mean ± standard deviation of AUC across reported evaluation splits; Δ is computed as full − no_time_gap. The full column is the diagnostic reference for this ablation and is distinct from the within-dataset tuned results in Table 1.

Dataset	full	no_time_gap	Δ	Interpretation
ASSIST2012	0.7618 ± 0.0039	0.7395 ± 0.0112	$+0.0223$	clear positive temporal effect
Junyi2015	0.7649 ± 0.0051	0.7513 ± 0.0084	$+0.0136$	positive effect in a clean-timestamp regime
MOOCRadar	0.7732 ± 0.0044	0.7561 ± 0.0062	$+0.0171$	positive effect in a clean-timestamp regime
ASSIST2017	0.7427 ± 0.0028	0.7410 ± 0.0049	$+0.0017$	marginal aggregate effect
XES3G5M	0.8149 ± 0.0078	0.8187 ± 0.0081	-0.0038	negative aggregate effect

6.4 TIME-GAP CONFIRMATION

Table 3 extends the aggregate temporal check to all five datasets. The positive differences on ASSIST2012, Junyi2015, and MOOCRadar are consistent with the timestamp-quality analysis, while ASSIST2017 shows only a marginal aggregate effect and XES3G5M shows a negative one. Thus, the table supports the claim that elapsed-time transport is useful when timestamp information is sufficiently informative, while also making the dataset dependence of the aggregate effect explicit. The diagnostic reference results complement, rather than replace, the gap-quality and mechanism analyses in Appendix B; the corresponding gap-stratified ASSIST2012 results are reported in Table 7.

7 DISCUSSION

Core contribution. The contribution is structural: BRIDGEFLOWKT casts KT as a continuous-time belief-flow system with explicit inter-event transport, bounded event updates, and query-conditioned prediction from the pre-event belief. Uncertainty is a transported state coordinate, and the theory and ablations test that mechanism directly.

Empirical summary. The evidence is mechanism-centered. BRIDGEFLOWKT wins clearly on ASSIST2012, attains the highest mean with a small margin on MOOCRadar, remains competitive on ASSIST2017, and stays in the top cluster on Junyi2015 and XES3G5M. Matched ablations show that residual query-state interaction and query-conditioned uncertainty matter most, with the clearest temporal gains on ASSIST2012 and dataset-dependent evidence elsewhere.

Scope and limitations. The continuous-time flow is intentionally simple—diagonal, monotone, and exactly solvable—which aids analysis and efficiency but cannot represent arbitrary continuous-time couplings. The uncertainty state is a structured chart rather than a full posterior, comparator coverage is partly constrained by dataset-specific annotations, and pooled AUC can obscure temporal effects in duplicate-heavy regimes. The main contribution is an interpretable continuous-time mean–uncertainty belief system with falsifiable mechanism tests.

8 CONCLUSION

We introduced BRIDGEFLOWKT, a continuous-time mean–uncertainty belief-flow system for knowledge tracing. It transports the learner state between events with exact exponential dynamics, updates it through bounded bridges, and predicts from a query-conditioned pre-event belief. Across the matched evaluations reported here, BRIDGEFLOWKT is competitive, residual query-state interaction is essential, and the uncertainty coordinate improves over a mean-only readout, with the clearest temporal evidence on ASSIST2012. The central design choice is to make time a state-transition operator rather than merely another observation-level feature, yielding a single framework in which temporal transport, evidence assimilation, and query-conditioned uncertainty can be analyzed separately. This separation turns the model from a collection of temporal heuristics into a structured and falsifiable account of how learner state evolves between responses. Overall, the results support BRIDGEFLOWKT as a coherent continuous-time architecture with theory and falsifiable mechanism tests.

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 (SIGIR '19)*, pp. 175–184, 2019.

Jiahao Chen, Zitao Liu, Shuyan Huang, Qiongqiong Liu, and Weiqi Luo. Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '23)*, volume 37, pp. 14196–14204, 2023.

Weihua Cheng, Hanwen Du, Chunxiao Li, Ersheng Ni, Liangdi Tan, Tianqi Xu, and Yongxin Ni. Uncertainty-aware Knowledge Tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '25)*, volume 39, pp. 27905–27913, 2025.

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 @ Scale (L@S '20)*, pp. 341–344, 2020.

Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298:103503, 2021.

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, 1994.

Chaoran Cui, Yumo Yao, Chunyun Zhang, Hebo Ma, Yuling Ma, Zhaochun Ren, Chen Zhang, and James Ko. DGEKT: A dual graph ensemble learning method for knowledge tracing. *ACM Transactions on Information Systems*, 42(3):1–24, 2024a.

Jiajun Cui, Hong Qian, Bo Jiang, and Wei Zhang. Leveraging pedagogical theories to understand student learning process with graph-based reasonable knowledge tracing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, pp. 502–513, 2024b.

Mingyu Feng, Neil Heffernan, and Kenneth 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.

Weibo Gao, Qi Liu, Rui Li, Yuze Zhao, Hao Wang, Linan Yue, Fangzhou Yao, and Zheng Zhang. Denoising programming knowledge tracing with a code graph-based tuning adaptor. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*, pp. 354–365, 2025.

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 (KDD '20)*, pp. 2330–2339, 2020.

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 (MM '21)*, pp. 367–375, 2021.

Liangliang He, Xiao Li, Pancheng Wang, Jintao Tang, and Ting Wang. MAN: Memory-augmented attentive networks for deep learning-based knowledge tracing. *ACM Transactions on Information Systems*, 42(1):1–22, 2023.

Shuyan Huang, Zitao Liu, Xiangyu Zhao, Weiqi Luo, and Jian Weng. Towards robust knowledge tracing models via k-sparse attention. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, pp. 2441–2445, 2023.

Shuyan Huang, Zitao Liu, Qiongqiong Liu, Jiahao Chen, and Yaying Huang. Dual-attentional time-aware fusion networks for knowledge tracing. *Information Fusion*, 126:103573, 2026.

Wonsung Lee, Jaeyoon Chun, Youngmin Lee, Kyoungsoo Park, and Sungrae Park. Contrastive learning for knowledge tracing. In *Proceedings of the ACM Web Conference 2022*, pp. 2330–2338, 2022.

Jiawei Li, Yuanfei Deng, Yixiu Qin, Shun Mao, and Yuncheng Jiang. Dual-Channel Adaptive Scale Hypergraph Encoders With Cross-View Contrastive Learning for Knowledge Tracing. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6752–6766, 2024.

Linqing Li, Zhifeng Wang, Joemon M Jose, and Xuri Ge. Llm supporting knowledge tracing leveraging global subject and student specific knowledge graphs. *Information Fusion*, pp. 103577, 2025.

Guimei Liu, Huijing Zhan, and Jung-jae Kim. Question difficulty consistent knowledge tracing. In *Proceedings of the ACM Web Conference 2024*, pp. 4239–4248, 2024.

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 (WWW '23)*, pp. 4178–4187, 2023a.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simpleKT: A Simple But Tough-to-Beat Baseline for Knowledge Tracing. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR '23)*, 2023b.

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:32958–32970, 2023c.

Ting Long, Li'ang Yin, Yi Chang, Wei Xia, and Yong Yu. Simulating Question-answering Correctness with a Conditional Diffusion. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*, pp. 5173–5182, 2025.

Haiping Ma, Yong Yang, Chuan Qin, Xiaoshan Yu, Shangshang Yang, Xingyi Zhang, and Hengshu Zhu. HD-KT: Advancing robust knowledge tracing via anomalous learning interaction detection. In *Proceedings of the ACM Web Conference 2024*, pp. 4479–4488, 2024.

Koki Nagatani, Qian Zhang, Masahiro Sato, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. Augmenting knowledge tracing by considering forgetting behavior. In *Proceedings of the World Wide Web Conference (WWW)*, pp. 3101–3107, 2019.

Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI '19)*, pp. 156–163, 2019.

Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM '19)*, pp. 384–389. International Educational Data Mining Society, 2019.

Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (LAK '13)*, pp. 117–124, 2013.

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 (NeurIPS 2015)*, pp. 505–513, 2015.

Yang Qin, Xinning Zhu, Xiaosheng Tang, Chunhong Zhang, Kunbao Wu, Fengjie Chang, Jianzhou Diao, and Zheng Hu. Interpretable Knowledge Tracing with Difficulty-Aware Attention and Selective State Space Model. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 316–325, 2025.

Shuanghong Shen, Qi Liu, Enhong Chen, Zhenya Huang, Wei Huang, Yu Yin, Yu Su, and Shijin Wang. Learning process-consistent knowledge tracing. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '21)*, pp. 1452–1460, 2021.

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 (SIGIR '22)*, pp. 427–437, 2022.

Shashank Sonkar, Andrew S. Lan, Andrew E. Waters, Phillip Grimaldi, and Richard G. Baraniuk. qDKT: Question-centric Deep Knowledge Tracing. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. International Educational Data Mining Society, 2020. URL https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_35.pdf.

Jianwen Sun, Fenghua Yu, Sannyuya Liu, Yawei Luo, Ruxia Liang, and Xiaoxuan Shen. Adversarial bootstrapped question representation learning for knowledge tracing. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, pp. 8016–8025, 2023.

Jianwen Sun, Fenghua Yu, Qian Wan, Qing Li, Sannyuya Liu, and Xiaoxuan Shen. Interpretable knowledge tracing with multiscale state representation. In *Proceedings of the ACM Web Conference 2024*, pp. 3265–3276, 2024.

Xinjie Sun, Kai Zhang, Qi Liu, Shuanghong Shen, Fei Wang, Yuxiang Guo, and Enhong Chen. Daskt: A dynamic affect simulation method for knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1714–1727, 2025. doi: 10.1109/TKDE.2025.3526584.

Hanshuang Tong, Zhen Wang, Yun Zhou, Shiwei Tong, Wenyuan Han, and Qi Liu. Introducing problem schema with hierarchical exercise graph for knowledge tracing. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*, pp. 405–415, 2022.

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 (WSDM '21)*, pp. 517–525, 2021.

Shanshan Wang, Fangzheng Yuan, Keyang Wang, Xun Yang, Xingyi Zhang, and Meng Wang. Dual-state personalized knowledge tracing with emotional incorporation. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2440–2455, 2025. doi: 10.1109/TKDE.2025.3538121.

Bihan Xu, Zhenya Huang, Jiayu Liu, Shuanghong Shen, Qi Liu, Enhong Chen, Jinze Wu, and Shijin Wang. Learning behavior-oriented knowledge tracing. In *Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '23)*, pp. 2789–2800, 2023.

Yang Yang, Jian Shen, Yanru Qu, Yunfei Liu, Kerong Wang, Yaoming Zhu, Weinan Zhang, and Yong Yu. GIKT: a graph-based interaction model for knowledge tracing. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 299–315. Springer, 2020.

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 (WWW '23)*, pp. 855–864, 2023.

Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, et al. MOOC-Radar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in MOOCs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, pp. 2924–2934, 2023.

Xiaoshan Yu, Chuan Qin, Dazhong Shen, Shangshang Yang, Haiping Ma, Hengshu Zhu, and Xingyi Zhang. RIGL: A unified reciprocal approach for tracing the independent and group learning processes. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, pp. 4047–4058, 2024.

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 (WWW '17)*, pp. 765–774, 2017.

A Extended Related Work

Foundations, benchmarks, and tooling. Knowledge tracing originated from interpretable latent mastery modeling in BKT (Corbett & Anderson, 1994) and was later reshaped by neural sequence prediction through DKT (Piech et al., 2015). The broader KT literature is now supported by benchmark-oriented surveys, shared evaluation toolchains, and widely used educational datasets including ASSIST-family corpora, MOOCRadar, and XES3G5M (Feng et al., 2009; Pardos et al., 2013; Yu et al., 2023; Liu et al., 2023c). This paper follows that benchmark tradition, but uses the empirical board mainly to test whether continuous-time transport and uncertainty-aware readout behave as predicted by the model design. Work on explainable intelligent tutoring systems also motivates the need for mechanisms whose behavior can be inspected rather than only scored by aggregate prediction metrics (Conati et al., 2021).

Neural sequence, memory, and attention KT. Early neural KT models replaced explicit probabilistic state updates with recurrent sequence encoders (Piech et al., 2015). Memory-augmented variants such as DKVMN and SKVMN use key-value memories to represent latent concepts and their evolving mastery (Zhang et al., 2017; Abdelrahman & Wang, 2019), while MAN combines memory and attention to handle skill switching (He et al., 2023). Attention-based KT models, including SAKT, SAINT, AKT, SimpleKT, and SparseKT, improve long-range dependency modeling and robust interaction selection (Pandey & Karypis, 2019; Choi et al., 2020; Ghosh et al., 2020; Liu et al., 2023b; Huang et al., 2023). Other neural improvements add adversarial training, auxiliary tasks, diagnostic transformer objectives, or contrastive learning (Guo et al., 2021; Liu et al., 2023a; Yin et al., 2023; Lee et al., 2022). BridgeFlowKT differs from this line by making the state transition itself a closed-form continuous-time operation rather than another learned discrete sequence block.

Time, forgetting, process, and behavioral evidence. Several KT models are explicitly motivated by the fact that student knowledge changes with time and learning process. DKT-Forget augments DKT with forgetting-related elapsed-time and practice-count features (Nagatani et al., 2019). HawkesKT models temporal cross-effects through point-process style excitation and decay (Wang et al., 2021). LPKT and LBKT incorporate learning-process consistency, answer time, elapsed time, and behavior features such as attempts or hints (Shen et al., 2021; Xu et al., 2023). Recent time-aware fusion models also emphasize elapsed-time information in prediction (Huang et al., 2026). Affect- or emotion-aware KT models expand the evidence source beyond correctness alone (Sun et al., 2025; Wang et al., 2025). These works support the importance of non-indexed temporal and behavioral information; BridgeFlowKT focuses on a narrower object, namely exact transport of the latent belief state before event assimilation.

Question, concept, difficulty, and graph structure. Many KT gains come from better modeling of questions, concepts, and their relations. Graph-based methods represent concept or exercise structure through graph neural networks, interaction graphs, hierarchical exercise graphs, reasonable knowledge-transition graphs, hypergraphs, or dual graph ensembles (Nakagawa et al., 2019; Yang et al., 2020; Tong et al., 2022; Cui et al., 2024b; Li et al., 2024; Cui et al., 2024a). Question-centric and difficulty-aware methods address the limitation of treating all questions under the same concept as interchangeable (Chen et al., 2023; Sonkar et al., 2020; Liu et al., 2024; Shen et al., 2022). Interpretable and multiscale models such as MIKT and ASIKT further emphasize diagnostic state structure and readable prediction mechanisms (Sun et al., 2024; Qin et al., 2025). Representation-learning methods such as ABQR improve question embeddings through graph augmentation and bootstrapped learning (Sun et al., 2023). BridgeFlowKT uses question and concept embeddings, but its main contribution is not a new item graph; it is the query-conditioned readout of a transported mean–uncertainty belief chart.

Robustness, new domains, and external knowledge. Recent KT work also studies robustness and broader educational scenarios. HD-KT detects anomalous learning interactions to reduce the effect of noisy behaviors (Ma et al., 2024). DISKT adapts KT ideas to programming knowledge tracking by using code-graph structure to reduce noise in programming submissions (Gao et al., 2025). Diffusion-style response simulation and reciprocal individual–group learning extend the range of modeling assumptions used in KT (Long et al., 2025; Yu et al., 2024). Large-language-model and knowledge-graph supported KT methods explore external semantic knowledge and cold-start settings

(Li et al., 2025). These directions are complementary to BRIDGEFLOWKT: they can potentially provide richer event features, while our contribution concerns how such evidence is assimilated into a continuous-time belief process.

Uncertainty-aware KT. UKT is the closest direct prior work for the uncertainty side of the model (Cheng et al., 2025). It introduces stochastic representations and uncertainty-aware matching, showing that uncertainty can be useful in neural KT. The distinction is the modeling object. UKT attaches uncertainty to representation and matching mechanisms, whereas BRIDGEFLOWKT treats uncertainty as a coordinate of the learner belief state, transports it continuously in time, updates it through bounded event bridges, and exposes it through query-conditioned scalar and interaction paths. This is why the experiments compare directly with UKT while also testing internal uncertainty masks.

B EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

B.1 EXPERIMENTAL PROTOCOL

All methods are trained under the same benchmark split protocol and the same basic optimization schedule unless an ablation explicitly changes one component. The empirical evaluation is AUC-centered: for each predictive result we report the mean and dispersion over evaluation splits, and we use timestamp-quality statistics and mechanism ablations to interpret when aggregate AUC is or is not informative.

The main predictive board covers five target datasets, including ASSIST-family datasets, MOOCRadar, and XES3G5M (Feng et al., 2009; Pardos et al., 2013; Yu et al., 2023; Liu et al., 2023c). Mechanism ablations are reported on ASSIST2012, ASSIST2017, and XES3G5M, which provide complementary sequence scales and timestamp regimes. These tables are used to isolate when elapsed-time transport and uncertainty-aware readout matter, rather than to argue that every component improves every dataset. Partially observed comparator coverage is marked explicitly.

Table 4: Mapping between theoretical claims and empirical tests. Each row links a formal claim to its corresponding empirical test and its clearest falsifying pattern.

Claim	Formal object	Empirical test	Falsifying pattern
Continuous time matters	Semigroup transport and gap-aliasing separation	full vs no_time_gap; timestamp-quality diagnostics	no_time_gap matches or beats full on timestamp-informative matched evaluations
Uncertainty contributes beyond mean state	Mean-equivalent state separation	full vs mu-only readout	mean-only remains close to the full model across matched ablations
Query-conditioned uncertainty matters	Interaction coordinates inside $U_q(\boldsymbol{\rho})$	no scalar uncertainty vs no interaction uncertainty	removing interaction uncertainty does not hurt more than removing the scalar path
Auxiliary losses regularize long-horizon behavior	Proper anytime BCE and flow-matching optimum	no_anytime, no_flow, no_consistency, weak_aux	removing matched auxiliary supervision does not reduce validation AUC
Distinct uncertainty modeling matters	Different uncertainty modeling object	matched UKT baseline under the same evaluation protocol	matched UKT matches or exceeds BRIDGEFLOWKT throughout the reported comparisons

B.2 DATASETS

The main board uses ASSIST2012[1], ASSIST2017[2], Junyi2015[3], MOOCRadar-C746997[4], and XES3G5M[5] under the benchmark split protocol used in this study. Table 5 summarizes the core preprocessing statistics: numbers of students, interactions, questions, and concepts. The ASSIST-family datasets, MOOCRadar, and XES3G5M are established KT benchmarks (Feng et al., 2009; Pardos et al., 2013; Yu et al., 2023; Liu et al., 2023c). The elapsed-time summary retained in Table 6 is the median non-zero inter-event gap, computed after excluding zero-gap ties.

Table 5: Dataset statistics from preprocessing metadata used in this study.

Dataset	Students	Interactions	Questions	Concepts
ASSIST2012	29,018	2,711,813	53,091	265
ASSIST2017	1,709	864,713	2,210	101
Junyi2015	247,606	25,925,987	817	40
MOOCRadar	1,577	100,066	550	265
XES3G5M	18,066	5,549,184	7,652	865

Because the temporal claim depends on the informativeness of timestamps rather than only their presence, Table 6 reports train-split zero-gap, duplicate-timestamp, and long-gap statistics for the same benchmark datasets together with the median non-zero inter-event gap. ASSIST2012, Junyi2015, and MOOCRadar are comparatively clean; ASSIST2017 is mostly clean but ultra-dense, leaving little elapsed-time dynamic range; XES3G5M is strongly duplicate-heavy. We therefore treat the first group as the clearest evidence for continuous-time behavior, ASSIST2017 as a mild-support regime, and XES3G5M as a stress-test regime rather than a decisive temporal benchmark.

Table 6: Timestamp-quality diagnostics. Percentages are ratios over adjacent within-sequence event pairs on the train split. Median non-zero gap is the median inter-event time difference after excluding zero-gap ties.

Dataset	Zero-gap	Duplicate ts	Long gap $>$ 30d	Median non-zero gap	Regime
ASSIST2012	0.54%	0.54%	0.61%	47s	clean
ASSIST2017	0.91%	0.90%	0.35%	12s	ultra-dense
Junyi2015	0.03%	0.03%	0.12%	12s	clean
MOOCRadar	0.001%	0.001%	0.08%	18s	clean
XES3G5M	51.63%	51.33%	0.47%	19m	duplicate-heavy

B.3 BASELINES

The baseline board includes classical and neural KT models such as DKT, DKVMN, ATKT, AKT, AT-DKT, SimpleKT, SparseKT, QIKT, qDKT, LPKT, LBKT, DKT-Forget, and UKT (Piech et al., 2015; Zhang et al., 2017; Guo et al., 2021; Ghosh et al., 2020; Liu et al., 2023a;b; Huang et al., 2023; Chen et al., 2023; Sonkar et al., 2020; Shen et al., 2021; Xu et al., 2023; Nagatani et al., 2019; Cheng et al., 2025). UKT is the key direct prior work and provides the closest comparison for the uncertainty component. Some cells are structurally unavailable rather than missing at random: under the preprocessing used here, LPKT requires elapsed-time and use-time features, whereas LBKT additionally requires hint and attempt-count features. MOOCRadar-C746997 and XES3G5M do not provide this full set of annotations, so the corresponding LPKT/LBKT entries are omitted; by contrast, the Junyi2015 entries reported here reflect compatible feature availability.

[1] https://sites.google.com/site/assistmentsdata/home/2012-13-school-data-with

[2] https://sites.google.com/view/assistmentsdatamining/dataset?authuser=0

[3] https://www.junyiacademy.org/

[4] https://github.com/THU-KEG/MOOC-Radar

[5] https://github.com/ai4ed/XES3G5M

B.4 METRICS

Reported predictive metrics are AUC for the main predictive board, time-gap ablations, uncertainty ablations, auxiliary-loss tuning, readout diagnostics, and interaction-head comparisons. We also include timestamp-quality statistics such as zero-gap rate, duplicate-timestamp rate, long-gap rate, and median non-zero gap to contextualize when elapsed-time transport should be expected to matter.

Statistical reporting. The tables report means and standard deviations across the reported evaluation splits. Small-sample comparisons are interpreted cautiously, and the strongest empirical evidence is therefore a repeated directional pattern across datasets and mechanism ablations rather than a single isolated p-value.

B.5 GAP-SENSITIVE INTERPRETATION

Timestamp-sensitive diagnostics are most informative when aggregate AUC is ambiguous. On short-gap-dominated or duplicate-heavy datasets, pooled ranking metrics can obscure whether elapsed-time transport is being exercised in meaningful separated histories. For that reason, the appendix reports timestamp-quality statistics, gap-stratified ablations, and mechanism ablations as context for the aggregate AUC results. Calibration-oriented uncertainty diagnostics are reported separately below.

From this perspective, ASSIST2012 provides the clearest aggregate confirmation, ASSIST2017 shows only a small aggregate gain, and XES3G5M mainly delimits where the temporal claim should remain conservative. The same logic applies to uncertainty: the evidence now shows a consistent gain from the transported uncertainty coordinate, with the clearest advantage coming from query-conditioned interaction paths and an additional contribution from scalar uncertainty features. The purpose of these diagnostics is therefore to localize where the continuous-time transport and query-conditioned uncertainty mechanisms are actually being exercised.

B.6 GAP-STRATIFIED TEMPORAL ANALYSIS

To examine whether the aggregate temporal comparison is driven by interactions for which elapsed time is informative, we further stratified the ASSIST2012 diagnostic evaluation by the physical gap from the preceding interaction. Each prediction event was assigned to exactly one gap interval. "Est. prop." is the approximate fraction of evaluation events in that interval, while each AUC entry is computed within the corresponding subset. The full and no_time_gap models use the same student-level partitions, preprocessing, and evaluation splits; the only intended difference is whether elapsed-time transport is active. Standard deviations are taken across the reported evaluation splits.

Table 7: Gap-stratified temporal comparison on ASSIST2012. The absolute Δ is full AUC minus no_time_gap AUC within each gap stratum.

Time-gap stratum	Est. prop.	full AUC	no_time_gap AUC	Absolute Δ
Extremely short ($<$ 15s)	$\sim 20\%$	0.7431 ± 0.0063	0.7373 ± 0.0078	$+0.0058$
Active memory (15s–5m)	$\sim 55\%$	0.7663 ± 0.0045	0.7426 ± 0.0086	$+0.0237$
Short delay (5m–24h)	$\sim 15\%$	0.7624 ± 0.0057	0.7317 ± 0.0121	$+0.0307$
Long-term ($>$ 24h)	$\sim 10\%$	0.7578 ± 0.0081	0.7196 ± 0.0145	$+0.0382$

The temporal margin grows monotonically across the four strata, from $+0.0058$ for extremely short gaps to $+0.0382$ for gaps longer than one day. The no_time_gap reference also decreases from 0.7373 to 0.7196 as the gap becomes longer, whereas the full model remains in the 0.7578–0.7663 range. This pattern is consistent with the proposed mechanism: when the interval is too short to induce much state transport, removing the flow has a smaller effect; when the interval is longer, the transported state provides information unavailable to a gap-agnostic model. Because the analysis is observational within one benchmark, it should be read as mechanism-localizing evidence rather than as a causal estimate of forgetting alone.

B.7 AUXILIARY-LOSS TUNING

Table 8: Auxiliary-loss tuning on ASSIST2012 in AUC. Entries are mean ± standard deviation across reported evaluation splits, and em dashes denote unavailable entries.

Variant	Validation AUC
full	0.7618 ± 0.0039
no_anytime	0.7490 ± 0.0051
anytime_005	0.7512 ± 0.0055
anytime_010	0.7565 ± 0.0052
short_anytime_010	0.7541 ± 0.0047
weak_aux	0.7528 ± 0.0055
no_aux	0.7482 ± 0.0038
residual_075	0.7645 ± 0.0054
residual_100	0.7668 ± 0.0039
residual_075_no_anytime	0.7505 ± 0.0056
residual_100_no_anytime	0.7527 ± 0.0055

Table 8 is primarily about recipe selection rather than a new scientific claim. On ASSIST2012, keeping anytime loss is beneficial: the full model exceeds no-anytime by about 0.0128, and increasing the residual scale to 1.0 while retaining anytime gives the strongest tuning result reported here at 0.7668. The residual_075 variant also improves over the full setting, while reduced-anytime variants such as anytime_010 and short-anytime_010 recover only part of the gain. Across the reported evaluation splits, no-aux and no-anytime are both clearly weaker than the matched anytime settings. Table 9 then shows that the top three reduced-anytime recipes remain below the transferred dynamics winner on all three clean datasets. This indicates that the preferred regime is not a flat low-anytime basin, but a matched combination of auxiliary supervision, residual scaling, and continuous-time dynamics.

Tuning-result narrative. The tuning results calibrate the training recipe rather than introducing a new scientific claim. They indicate that the weight 0.2 provides useful multi-horizon supervision on ASSIST2012, and that stronger residual scaling works best when coupled with that supervision. Across the clean-dataset confirmations, reduced-anytime variants remain consistently below the transferred dynamics winner, which means the tuning story is not about escaping an over-regularized objective by removing auxiliaries. Instead, the more consequential recipe choice is to preserve matched auxiliary constraints while calibrating the continuous-time dynamics and residual scale.

B.8 CLEAN-TIMESTAMP CONFIRMATION AND DYNAMICS TRANSFER

Table 9: Within-dataset confirmation on the timestamp-clean subset. The first three rows are the top three low-anytime recipes confirmed on ASSIST2012, Junyi2015, and MOOCRadar. The last row transfers the ASSIST2012 dynamics winner to the same clean subset.

Recipe	ASSIST2012	Junyi2015	MOOCRadar
top-1 low-anytime	0.7565	0.7421	0.7667
top-2 low-anytime	0.7541	0.7405	0.7651
top-3 low-anytime	0.7527	0.7396	0.7641
ASSIST2012-tuned dynamics transfer	0.7674	0.7811	0.7728

Table 9 adds two clarifications that are hard to see from the one-dataset tuning table alone. First, the top three low-anytime recipes now form a clear descending ladder and remain below the transferred dynamics winner on all three clean datasets, which indicates a persistent benefit from keeping matched auxiliary supervision. Second, the ASSIST2012 dynamics-tuned winner improves ASSIST2012 by about $+0.0109$ over the best clean-confirmed low-anytime recipe and also remains strongest on Junyi2015 and MOOCRadar, so the preferred recipe is better viewed as a coordinated tuning of dynamics and auxiliary weighting rather than as a brittle one-dataset accident.

B.9 READOUT DIAGNOSTICS AND STRUCTURE CHECKS

Table 10: Readout diagnostics on ASSIST2012 in AUC. These variants test whether the gains come from dynamic state-conditioned computation rather than from static question bias or static embeddings. The interaction-only head rows use aligned head-selection confirmations where noted.

Variant	Question answered	Validation AUC
no_concept_logit	Does the concept-side main term add unique value beyond the residual state-conditioned readout?	0.7627
state_readout_only	Can the transported state summary alone retain most of the signal?	0.7606
no_question_bias	Does removing question bias matter once the dynamic readout remains active?	0.7592
interaction_only	Can interaction features alone remain competitive once static shortcuts are suppressed?	0.7588
no_energy	Does the main readout signal survive without the energy regularizer?	0.7587
interaction_mlp_only_r96	Does a structured nonlinear interaction head remain strong when isolated?	0.7561
linear_residual	How much of the signal survives if the residual head is forced to be linear?	0.7315
static_embeddings_only	How much can the model gain from static question and concept embeddings alone?	0.6915
question_bias_only	Can simple item bias explain the observed improvements?	0.6564
summary_only	Can the pooled state summary alone explain the score?	0.6468

This table guards against a shallow explanation of the gains. In these ASSIST2012 diagnostics, removing the concept-side main term or question bias barely changes AUC, both the state-only and interaction-only dynamic paths remain competitive within this diagnostic set, and removing the energy penalty leaves the main ordering largely intact. By contrast, forcing the residual to be linear already drops noticeably, and static embeddings, question bias, or summary-only readouts collapse sharply. The useful signal therefore comes from dynamic state-conditioned computation rather than from static item features, memorized question difficulty, or a trivial pooled state summary.

Readout-result narrative. The readout diagnostics form an attribution result. The strongest signal is that the concept-side main term is not the main driver: no_concept_logit reaches 0.7627 and is among the strongest results in these diagnostics. state_readout_only (0.7606), no_question_bias (0.7592), and interaction_only (0.7588) each remain competitive, whereas linear_residual, static_embeddings_only, question_bias_only, and summary_only degrade much more sharply. The extended eliminations therefore reinforce the view that BridgeFlowKT's signal is dynamic and state-conditioned rather than static or shallow. The head-selection comparison in Table 11 then shows that the MLP, FiLM, and FiLM-gain interaction heads cluster together, while operator and PP-mix lag behind. Taken together, these results support the claim that BRIDGEFLOWKT's predictive strength comes from dynamic state-conditioned pathways rather than from static shortcuts or a single hand-picked interaction parameterization.

B.10 INTERACTION-HEAD FAMILY COMPARISON

Table 11: Interaction-head family comparison in AUC.

Head family	ASSIST2012	XES3G5M
FiLM-gain	0.7575	0.8089
FiLM	0.7569	0.8075
MLP	0.7561	0.8087
PP-mix	0.7306	0.8069
Operator	0.7262	0.8075

On ASSIST2012, the FiLM-gain, FiLM, and MLP heads form a tight top group, whereas operator and PP-mix are clearly weaker. On XES3G5M the stronger heads cluster much more tightly, which is consistent with that dataset being less informative for fine-grained temporal mechanism selection. These results indicate that BRIDGEFLOWKT does not depend on a single idiosyncratic interaction parameterization.

B.11 UNCERTAINTY CALIBRATION AND EDUCATIONAL-FACTOR ASSOCIATIONS

The AUC ablations test whether the uncertainty coordinate improves prediction, but they do not by themselves show whether its output behaves like a useful confidence signal. We therefore add two complementary diagnostics on ASSIST2012. First, Table 12 compares the full model with uncertainty-reduced variants using negative log-likelihood (NLL), Brier score, and expected calibration error (ECE), in addition to AUC. NLL evaluates the quality of predicted probabilities under log loss, the Brier score measures squared probability error, and ECE summarizes the discrepancy between confidence and empirical accuracy across probability bins. All three variants use the no-anytime setting so that the comparison matches the direct uncertainty ablations in Table 2; lower is better for NLL, Brier score, and ECE.

Table 12: Calibration-oriented diagnostics on the ASSIST2012 diagnostic reference runs. Entries are mean $\pm$ standard deviation across the reported evaluation splits; all variants use the no-anytime setting.

Variant (no-anytime)	AUC	NLL	Brier score	ECE
full	0.7490 ± 0.0051	0.5314 ± 0.0028	0.1782 ± 0.0019	0.0187 ± 0.0022
no_interaction_uncertainty	0.7396 ± 0.0054	0.5429 ± 0.0035	0.1837 ± 0.0024	0.0351 ± 0.0041
mu_only_readout	0.7383 ± 0.0049	0.5487 ± 0.0042	0.1861 ± 0.0029	0.0413 ± 0.0038

The full model is better on every reported metric. Relative to `no_interaction_uncertainty`, it improves AUC by 0.0094, NLL by 0.0115, Brier score by 0.0055, and ECE by 0.0164. Relative to `mu_only_readout`, the corresponding differences are 0.0107, 0.0173, 0.0079, and 0.0226. Thus, the uncertainty pathway contributes not only to ranking quality but also to the quality of the predicted probabilities in this diagnostic setting. In particular, the lower ECE indicates that the full model's confidence is closer to its empirical accuracy than the two uncertainty-reduced readouts. These results support a calibration-relevant interpretation of the channel, while not claiming that ρ is a fully Bayesian posterior variance.

Second, we test whether the learned uncertainty scale is related to observable learning-process factors. For every eligible ASSIST2012 interaction, we compute the transported scale $\sigma(\rho)$ and the operational factor measured from the prefix available before that interaction. The analysis is interaction-level because each response event supplies one state–factor pair; however, repeated events from the same student are not treated as independent for inference. Table 13 reports Spearman rank correlations and student-clustered bootstrap significance estimates.

Table 13: Associations between the learned uncertainty scale $\sigma(\rho)$ and educational factors on ASSIST2012. Correlations use interaction-level observations, with student-clustered bootstrap inference.

Educational factor	Operational metric	Spearman r	p-value	Interpretation
Practice volume	$\log(1 + \text{prefix history length})$	-0.428	< 0.001	uncertainty decreases as observed practice accumulates
Elapsed time gap	physical inter-event delay	$+0.385$	< 0.001	longer idle intervals are associated with higher learner-state ambiguity
Concept novelty	inverse prior concept exposure frequency	$+0.451$	< 0.001	rarer or less familiar concepts have higher uncertainty
Response volatility	moving variance of recent correctness	$+0.312$	< 0.001	inconsistent recent responses correspond to higher uncertainty

The four associations have the directions expected from the model's intended interpretation. The negative practice-volume correlation indicates lower uncertainty after more observed practice, while the positive gap correlation is consistent with uncertainty increasing after a longer period without an interaction. The novelty and volatility correlations further show that the channel responds to concept familiarity and response consistency rather than only to sequence length. The magnitudes are moderate rather than deterministic, which is appropriate for a latent state influenced by multiple factors. We therefore view Table 13 as evidence that $\sigma(\rho)$ is educationally structured and calibration-relevant, not as external validation of a probabilistically defined learner uncertainty.

C Full Proofs

Throughout, we use two structural facts from the main text: between events the centered state evolves by the linear semigroup $S_\tau = e^{-\tau\Lambda}$, and each event update is a bounded jump followed by the nonexpansive clipping map Π_c.

C.1 Proof of Theorem 1

Let $\tilde{\mathbf{z}}(t) = \mathbf{z}(t) - \mathbf{z}_{\mathrm{eq}}$. Equation (12) becomes

$$\frac{\mathrm{d}\tilde{\mathbf{z}}(t)}{\mathrm{d}t} = -\Lambda\tilde{\mathbf{z}}(t), \qquad \tilde{\mathbf{z}}(0) = \mathbf{z}_0 - \mathbf{z}_{\mathrm{eq}}. \tag{44}$$

Because this is a linear autonomous ODE in finite dimension, its unique solution is given by the matrix exponential,

$$\tilde{\mathbf{z}}(t) = e^{-t\Lambda}\tilde{\mathbf{z}}(0) = S_t\tilde{\mathbf{z}}(0). \tag{45}$$

Re-centering yields

$$\Phi_\tau(\mathbf{z}) = \mathbf{z}_{\mathrm{eq}} + S_\tau(\mathbf{z} - \mathbf{z}_{\mathrm{eq}}), \tag{46}$$

which is Equation (13). To prove the semigroup law, compute

$$\begin{aligned}\Phi_{\tau+s}(\mathbf{z}) &= \mathbf{z}_{\mathrm{eq}} + S_{\tau+s}(\mathbf{z} - \mathbf{z}_{\mathrm{eq}}) \\ &= \mathbf{z}_{\mathrm{eq}} + S_\tau S_s(\mathbf{z} - \mathbf{z}_{\mathrm{eq}}) \\ &= \mathbf{z}_{\mathrm{eq}} + S_\tau\big(\Phi_s(\mathbf{z}) - \mathbf{z}_{\mathrm{eq}}\big) \\ &= \Phi_\tau\big(\Phi_s(\mathbf{z})\big).\end{aligned} \tag{47}$$

Also $\Phi_0 = \mathrm{Id}$ because $S_0 = e^{0\Lambda} = I$. Finally,

$$\|\Phi_\tau(\mathbf{z}) - \Phi_\tau(\mathbf{z}')\|_2 = \|S_\tau(\mathbf{z} - \mathbf{z}')\|_2 \le \|S_\tau\|_2\|\mathbf{z} - \mathbf{z}'\|_2. \tag{48}$$

Since S_τ is diagonal with entries $e^{-\lambda_m\tau}$,

$$\|S_\tau\|_2 = \max_m e^{-\lambda_m\tau} = e^{-\lambda_{\min}\tau}, \tag{49}$$

which proves the contraction inequality.

C.2 Proof of Theorem 2

Without projection, the centered recurrence is

$$\tilde{\mathbf{z}}_i^- = A_i \tilde{\mathbf{z}}_{i-1}^+, \qquad \tilde{\mathbf{z}}_i^+ = \tilde{\mathbf{z}}_i^- + \Delta \mathbf{z}_i, \qquad A_i = e^{-\Delta t_i \Lambda}. \tag{50}$$

Let $T_i = \sum_{r=1}^{i} \Delta t_r$, $P_i = e^{-T_i \Lambda}$, and $P_0 = I$. Because every A_i is a function of the same diagonal operator Λ, the matrices commute and satisfy

$$A_i A_{i-1} \cdots A_{j+1} = e^{-(T_i - T_j)\Lambda} = P_i P_j^{-1}. \tag{51}$$

Unrolling the recurrence gives

$$\begin{aligned} \tilde{\mathbf{z}}_i^- &= A_i A_{i-1} \cdots A_1 \tilde{\mathbf{z}}_0^+ + \sum_{j<i} A_i A_{i-1} \cdots A_{j+1} \Delta \mathbf{z}_j \\ &= P_i \tilde{\mathbf{z}}_0^+ + \sum_{j<i} P_i P_j^{-1} \Delta \mathbf{z}_j \\ &= P_i \left(\tilde{\mathbf{z}}_0^+ + \sum_{j<i} P_j^{-1} \Delta \mathbf{z}_j \right). \end{aligned} \tag{52}$$

For the post-event state,

$$\begin{aligned} \tilde{\mathbf{z}}_i^+ &= \tilde{\mathbf{z}}_i^- + \Delta \mathbf{z}_i \\ &= P_i \left(\tilde{\mathbf{z}}_0^+ + \sum_{j<i} P_j^{-1} \Delta \mathbf{z}_j \right) + P_i P_i^{-1} \Delta \mathbf{z}_i \\ &= P_i \left(\tilde{\mathbf{z}}_0^+ + \sum_{j \le i} P_j^{-1} \Delta \mathbf{z}_j \right). \end{aligned} \tag{53}$$

This is exactly Equation (18). Hence the scan uses only exponentials and cumulative sums and requires no numerical ODE integration.

C.3 Proof of Proposition 1

Let $d_i = \|\mathbf{z}_i^+ - \bar{\mathbf{z}}_i^+\|_2$ and $\delta_i = \Delta \mathbf{z}_i - \Delta \bar{\mathbf{z}}_i$. Because the two paths use the same time gaps, they share the same transport map $\Phi_{\Delta t_i}$. By nonexpansiveness of Π_c and Theorem 1,

$$\begin{aligned} d_i &= \left\| \Pi_c\big(\Phi_{\Delta t_i}(\mathbf{z}_{i-1}^+) + \Delta \mathbf{z}_i\big) - \Pi_c\big(\Phi_{\Delta t_i}(\bar{\mathbf{z}}_{i-1}^+) + \Delta \bar{\mathbf{z}}_i\big) \right\|_2 \\ &\le \left\| \Phi_{\Delta t_i}(\mathbf{z}_{i-1}^+) - \Phi_{\Delta t_i}(\bar{\mathbf{z}}_{i-1}^+) + \delta_i \right\|_2 \\ &\le \left\| \Phi_{\Delta t_i}(\mathbf{z}_{i-1}^+) - \Phi_{\Delta t_i}(\bar{\mathbf{z}}_{i-1}^+) \right\|_2 + \|\delta_i\|_2 \\ &\le e^{-\lambda_{\min} \Delta t_i} d_{i-1} + \|\delta_i\|_2. \end{aligned} \tag{54}$$

Iterating this discrete Grönwall recursion yields

$$d_n \le e^{-\lambda_{\min} T_n} d_0 + \sum_{j=1}^{n} e^{-\lambda_{\min}(T_n - T_j)} \|\Delta \mathbf{z}_j - \Delta \bar{\mathbf{z}}_j\|_2, \tag{55}$$

which is Equation (34).

C.4 Proof of Proposition 2

The post-event claim is immediate because Π_c projects onto the compact box $\mathcal{Z}_c$. For a pre-event state,

$$\mathbf{z}_i^- = \mathbf{z}_{\mathrm{eq}} + S_{\Delta t_i}(\mathbf{z}_{i-1}^+ - \mathbf{z}_{\mathrm{eq}}). \tag{56}$$

Write this coordinatewise. For each coordinate m,

$$(\mathbf{z}_i^-)_m = (1 - s_{i,m})(\mathbf{z}_{\mathrm{eq}})_m + s_{i,m}(\mathbf{z}_{i-1}^+)_m, \qquad s_{i,m} = e^{-\lambda_m \Delta t_i} \in [0, 1]. \tag{57}$$

Hence every coordinate of $\mathbf{z}_i^-$ lies in the convex hull of the corresponding coordinates of $\mathbf{z}_{i-1}^+$ and $\mathbf{z}_{\mathrm{eq}}$. Since $\mathbf{z}_{i-1}^+ \in \mathcal{Z}_c$, the state $\mathbf{z}_i^-$ lies in the coordinatewise compact hull of $\mathcal{Z}_c \cup \{\mathbf{z}_{\mathrm{eq}}\}$, equivalently the axis-aligned compact box obtained by adjoining $\mathbf{z}_{\mathrm{eq}}$ to the coordinate ranges of $\mathcal{Z}_c$. Continuity of Γ_θ then implies boundedness on this reachable set.

C.5 Proof of Theorem 3

Let $\mathbf{z}(\tau) = \Phi_\tau(\mathbf{z}^+)$ and $r(\tau) = \log(1+\tau)$, so $g_q(\tau) = \Gamma_\theta(\mathbf{z}(\tau), q, \tau)$ with explicit gap coordinate $r(\tau)$. By Equation (13),

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\Phi_\tau(\mathbf{z}^+) = -\Lambda S_\tau(\mathbf{z}^+ - \mathbf{z}_{\mathrm{eq}}). \tag{58}$$

Since τ enters through both $\mathbf{z}(\tau)$ and $r(\tau)$, the chain rule gives

$$g_q'(\tau) = \nabla_{\mathbf{z}}\Gamma_\theta(\Phi_\tau(\mathbf{z}^+), q, \tau)^\top \frac{\mathrm{d}\Phi_\tau(\mathbf{z}^+)}{\mathrm{d}\tau} + \frac{\partial_r \Gamma_\theta(\Phi_\tau(\mathbf{z}^+), q, \tau)}{1+\tau}, \tag{59}$$

which is Equation (35). The integral bound then follows from the fundamental theorem of calculus:

$$g_q(\tau_2) - g_q(\tau_1) = \int_{\tau_1}^{\tau_2} g_q'(s)\,\mathrm{d}s. \tag{60}$$

Taking absolute values and substituting the derivative formula gives Equation (36).

C.6 Proof of Theorem 4

A gap-agnostic predictor depends only on the ordered question-response history together with non-temporal shortcuts. Once these variables are fixed, changing timestamps alone leaves its input unchanged, so its output is constant across all admissible gaps for that history class.

For BridgeFlowKT, fix a post-event state $\mathbf{z}^+$ and query q, and consider the temporal response $g_q(\tau)$ from Theorem 3. If g_q' is not identically zero on an interval I, then g_q is not constant on I. Hence there exist $\tau_1, \tau_2 \in I$ such that $g_q(\tau_1) \neq g_q(\tau_2)$. Because g_q is continuous, this strict inequality persists on small neighborhoods $U \ni \tau_1$ and $V \ni \tau_2$, so $g_q(u) \neq g_q(v)$ for every $(u, v) \in U \times V$. The set $U \times V \subset I^2$ has positive Lebesgue measure. Therefore BridgeFlowKT separates a positive-measure set of gap pairs on which every gap-agnostic predictor remains constant.

C.7 Proof of Theorem 5

If a readout factors through $\pi_\mu(\mathbf{z}) = \boldsymbol{\mu}$, then two mean-equivalent states $\mathbf{z} = (\boldsymbol{\mu}, \boldsymbol{\rho})$ and $\mathbf{z}' = (\boldsymbol{\mu}', \boldsymbol{\rho}')$ induce exactly the same input and hence exactly the same output. This proves the first claim.

For the full chart, all mean-dependent coordinates coincide because $\boldsymbol{\mu} = \boldsymbol{\mu}'$. Any difference must therefore lie in the uncertainty-chart coordinates $U_q(\boldsymbol{\rho})$ or in the uncertainty-dependent part of the state summary. In particular, Equation (21) gives

$$U_q(\boldsymbol{\rho}) - U_q(\boldsymbol{\rho}') = A_q^\sigma(\boldsymbol{\sigma}(\boldsymbol{\rho}) - \boldsymbol{\sigma}(\boldsymbol{\rho}')), \tag{61}$$

with A_q^σ as in Equation (37). Hence $A_q^\sigma(\boldsymbol{\sigma}(\boldsymbol{\rho}) - \boldsymbol{\sigma}(\boldsymbol{\rho}')) \neq \mathbf{0}$ is a sufficient feature-level separation in the full chart. If the downstream readout has nonzero local derivative in that direction, a first-order Taylor expansion yields local logit separation.

For the sensitivity formula, differentiate the uncertainty-dependent coordinates. Since $U_q(\boldsymbol{\rho}) = A_q^\sigma \boldsymbol{\sigma}(\boldsymbol{\rho})$,

$$D_{\boldsymbol{\rho}} U_q(\boldsymbol{\rho}) = A_q^\sigma \mathrm{diag}(\mathrm{softplus}'(\boldsymbol{\rho})). \tag{62}$$

The state-summary term contributes $D_{\boldsymbol{\rho}} H_s(\mathbf{z})$, whose nonzero part comes from the $\boldsymbol{\rho}$-summary block. Since the concept scaffold and question bias do not depend on $\boldsymbol{\rho}$, applying the chain rule only to the residual readout term gives Equation (38).

C.8 Proof of Proposition 3

Write the first layer of f_θ as $W_1\psi + b_1$. For a masked model, the first pre-activation is

$$W_1(M\psi) + b_1 = (W_1 M)\psi + b_1. \tag{63}$$

Therefore a full model can reproduce the same function by replacing the first-layer weight matrix with $\bar{W}_1 = W_1 M$ and leaving all subsequent layers unchanged. Equivalently, it sets to zero the

incoming weights attached to the masked coordinates. Hence every function in $\mathcal{M}_M$ is realized by some parameter choice in $\mathcal{M}_{\text{full}}$, proving

$$\mathcal{M}_M \subseteq \mathcal{M}_{\text{full}}. \tag{64}$$

The empirical-risk inequality in Equation (39) follows immediately because the infimum over a superset cannot exceed the infimum over a subset.

C.9 Proof of Proposition 4

Condition on $(\mathcal{F}_i^+, q_j, \tau)$ and write $p^\star = \mathbb{P}(Y_j = 1 \mid \mathcal{F}_i^+, q_j, \tau)$. Then Y_j is Bernoulli with mean $p^\star$, so for a prediction $p \in (0,1)$,

$$R(p) = -p^\star \log p - (1-p^\star)\log(1-p). \tag{65}$$

Its derivatives are

$$R'(p) = -\frac{p^\star}{p} + \frac{1-p^\star}{1-p}, \qquad R''(p) = \frac{p^\star}{p^2} + \frac{1-p^\star}{(1-p)^2} > 0. \tag{66}$$

Thus R is strictly convex on $(0,1)$. If $p^\star \in (0,1)$, solving $R'(p) = 0$ gives $p = p^\star$, the unique interior minimizer. If $p^\star \in \{0,1\}$, the usual extended log-loss convention gives the unique boundary minimizer $p = p^\star$.

C.10 Proof of Proposition 5

Let $\langle a, b\rangle_G = a^\top G b$ and $m(X_i) = \mathbb{E}[\boldsymbol{\xi}_i \mid X_i]$. For any square-integrable v,

$$\|v(X_i) - \boldsymbol{\xi}_i\|_G^2 = \|v(X_i) - m(X_i)\|_G^2 + \|\boldsymbol{\xi}_i - m(X_i)\|_G^2 + 2\langle v(X_i) - m(X_i), m(X_i) - \boldsymbol{\xi}_i\rangle_G. \tag{67}$$

Taking conditional expectation given X_i removes the cross term because $\mathbb{E}[\boldsymbol{\xi}_i - m(X_i) \mid X_i] = 0$. Therefore

$$\mathbb{E}[\|v(X_i) - \boldsymbol{\xi}_i\|_G^2 \mid X_i] = \|v(X_i) - m(X_i)\|_G^2 + \mathbb{E}[\|\boldsymbol{\xi}_i - m(X_i)\|_G^2 \mid X_i]. \tag{68}$$

The second term is independent of v, whereas the first is nonnegative and vanishes if and only if $v(X_i) = m(X_i)$ almost surely. Taking expectation proves that the unique minimizer is

$$v^\star(X_i) = m(X_i) = \mathbb{E}[\boldsymbol{\xi}_i \mid X_i]. \tag{69}$$

In the deterministic bridge used by BridgeFlowKT, $\boldsymbol{\xi}_i$ is already determined by $\mathbf{e}_i$, so the conditional mean reduces to $\boldsymbol{\xi}_i$ itself.

C.11 Proof of Proposition 6

Set $a_i = \|\Delta\mathbf{z}_i\|_G / \sqrt{w_i}$ and $b_i = \sqrt{w_i}$. Then $a_i b_i = \|\Delta\mathbf{z}_i\|_G$, so Cauchy–Schwarz gives

$$\begin{aligned}\sum_{i\in H} \|\Delta\mathbf{z}_i\|_G &= \sum_{i\in H} a_i b_i \\ &\le \left(\sum_{i\in H} a_i^2\right)^{1/2} \left(\sum_{i\in H} b_i^2\right)^{1/2} \\ &= \left(\sum_{i\in H} \frac{\|\Delta\mathbf{z}_i\|_G^2}{w_i}\right)^{1/2} \left(\sum_{i\in H} w_i\right)^{1/2}.\end{aligned} \tag{70}$$

This is exactly Equation (43).

D Additional Derivations

D.1 Predictive-mean rollout

Let $\Delta\mathbf{z}(q, y, \tau)$ denote the deterministic bridge displacement under a candidate response $y \in \{0,1\}$. Given $\mathbf{z}^-$ and predicted correctness probability p, set $\Delta\mathbf{z}_y = \Delta\mathbf{z}(q, y, \tau)$. When clipping is inactive, the one-step predictive mean satisfies

$$\mathbb{E}[\mathbf{z}^+ \mid \mathbf{z}^-, q, \tau] = \mathbf{z}^- + p\Delta\mathbf{z}_1 + (1-p)\Delta\mathbf{z}_0. \tag{71}$$

With clipping active, the exact conditional mean is $p\,\Pi_c(\mathbf{z}^- + \Delta\mathbf{z}_1) + (1-p)\,\Pi_c(\mathbf{z}^- + \Delta\mathbf{z}_0)$; applying Π_c after the averaged displacement is only a projected approximation. This identity is linear only in the bridge displacement; it does not imply $\mathbb{E}[\mathrm{sigmoid}(\Gamma_\theta(\mathbf{z}, q, \tau))] = \mathrm{sigmoid}(\Gamma_\theta(\mathbb{E}[\mathbf{z}], q, \tau))$ for nonlinear readouts.

E Implementation Details

This section records the shared-reference configuration used for mechanism ablations and cross-dataset coverage checks. Within-dataset tuning confirmations modify the loss weights or dynamics around this reference, as summarized in Tables 8 and 9, rather than implying a single universal optimum.

Following common conference appendix practice, we also record the local machine and software stack used for the BridgeFlowKT experiments reported here, including hardware, operating system, core deep-learning framework versions, and mixed-precision mode. We report only the components needed to understand or reproduce the BridgeFlowKT experiments, and omit unrelated framework-internal details.

Table 14: Local runtime environment and shared-reference BridgeFlowKT configuration used in ablations and coverage checks.

Component	Value
Operating system	Ubuntu 20.04.3 LTS (Linux 5.15.0-139-generic)
CPU	2× Intel Xeon Platinum 8350C (128 logical CPUs total)
GPU	4× NVIDIA GeForce RTX 4090 (24 GB each)
Driver / CUDA / cuDNN	535.183.01 / 12.1 / 9.1
Python / PyTorch	3.11.4 / 2.5.1+cu121
Training framework	Direct PyTorch implementation with custom training loops and CUDA autocast
Auxiliary libraries	NumPy 2.3.2, pandas 2.3.2, scikit-learn 1.7.1
Mixed precision	`precision=auto`; bf16 on this machine; GradScaler enabled only for fp16
Approx. time per batch	ASSIST2017: 13.6 ms/batch; ASSIST2012: 14.9 ms/batch; Junyi2015: 16.1 ms/batch; MOOCRadar: 26.3 ms/batch; XES3G5M: 24.6 ms/batch
State dimension d	96
Hidden dimension	256
Dropout	0.2
Optimizer	Adam, learning rate 5×10^{-4}, weight decay 10^{-4}
Train / eval batch size	32 / 64
Precision	auto
Epochs / early stopping	200 / 20
Max elapsed time	720 hours
State clip (c_μ, c_ρ)	6.0 / 6.0
Jump scales $(\delta_\mu, \delta_\rho)$	0.5 / 0.25
Initial decays $(\lambda_{\mu,0}, \lambda_{\rho,0})$	0.005 / 0.002
Initial uncertainty coordinate	-1.0
Residual scale γ	0.5
Main readout / uncertainty mode	mlp / full
Anytime offsets	[1,2,4,8,16]
Consistency offsets	[[1,4],[2,8],[4,16]]
$(\lambda_{\mathrm{any}}, \lambda_{\mathrm{cons}}, \lambda_{\mathrm{flow}}, \lambda_{\mathrm{energy}})$	$(0.2, 0.05, 0.05, 10^{-4})$
Split protocol	five-way benchmark split protocol
Sequence preprocessing	max sequence length 200, min sequence length 3