Staged Latent Chain-of-Thought for Sequential Counterfactual Prediction with Universal Holdout Calibration

Anonymous authors
Paper under double-blind review

Abstract

In personalized marketing and medicine, interventions shape both immediate outcomes and the states that influence future responses. Evaluating dynamic policies therefore requires predicting how state and outcome trajectories evolve under alternative intervention strategies. Long-term randomized holdouts are already used in practice, particularly in personalized marketing. Simply pooling these observations with intervention histories does not explicitly use randomized policy contrasts to supervise predicted trajectories.

We develop Staged Latent Chain-of-Thought with Universal Holdout Calibration (UHC), integrating task-aligned latent supervision with experimental calibration of generated trajectories. UHC transforms randomized comparisons within pretreatment-defined subgroups into time-specific causal anchors for state and outcome trajectories, complemented by within-arm distributional constraints. Staged Latent CoT supervises successive latent passes with observed state and outcome targets and aligns contrasts between the final rollouts with randomized subgroup anchors.

On a CVSim-based benchmark, the complete method reduces effect RMSE by 31.87% and policy-outcome RMSE by 23.07% relative to a G-Transformer trained on pooled treatment and holdout data. With architecture, parameter count, and inference computation held fixed, its advantage over unstaged latent chain-of-thought is significant on both metrics in all three training seeds after multiplicity correction. A separate matched ablation shows that state-then-outcome supervision also outperforms uniform and reversed schedules. The combined framework also improves Causal Transformer over its pooled baseline. This work introduces a framework for sequential counterfactual prediction that structures latent computation around states, outcomes, and effects while grounding generated trajectories in randomized holdout evidence.

1 Introduction

An intervention changes not only the next outcome, but also the conditions under which the next decision is made. A marketing action changes customer engagement and behavior used to select subsequent campaigns; a medical treatment changes the clinical state that guides subsequent care. Evaluating a dynamic policy therefore requires predicting a sequence of interactions between interventions, evolving states, and outcomes under alternative intervention strategies.

Long-term randomized control data already arise in practice. Marketing Universal Control Groups (UCGs) contain randomly selected users excluded from the intervention program across campaigns (Shectman, 2021); trials such as SPRINT randomize patients between standard and intensified treatment policies (The SPRINT Research Group, 2015). These data provide complementary training signals: observed trajectories in each arm and randomized between-arm comparisons. Simply pooling the observations supports trajectory prediction but does not explicitly use the randomized contrasts to supervise predicted policy effects.

Recent methods, including Causal Transformer (Melnychuk et al., 2022), G-Transformer (Xiong et al., 2024), and Mamba-CDSP (Wang et al., 2025), have advanced sequential counterfactual pre-

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.

diction from observational histories. Recurrent latent computation offers a way to refine predictions through repeated processing, with demonstrated benefits for structured-data prediction (Dudley & Oymak, 2026). For sequential counterfactual prediction, state predictions shape subsequent outcomes and policy decisions, and errors propagate into the final policy-effect estimate. This structure motivates assigning predictive roles to successive latent passes: learning how states evolve, predicting outcomes, and estimating effects from the resulting trajectories.

We introduce **Staged Latent Chain-of-Thought with Universal Holdout Calibration**, coupling task-aligned latent supervision with experimental calibration of generated trajectories. An early latent pass receives state-prediction supervision, a subsequent pass receives outcome-prediction supervision, and the final pass generates joint state–outcome predictions from which policy effects are computed. Successive stages communicate through latent representations and learn from observed trajectories. Universal Holdout Calibration (UHC) transforms randomized long-term holdouts into trajectory-level experimental supervision. Within pretreatment-defined subgroups, randomized comparisons supply time-specific causal anchors for state and outcome trajectories without requiring all response determinants to be observed. These anchors, together with within-arm distributional targets, constrain generated trajectories throughout follow-up. The framework thus connects the supervision of intermediate latent computation with experimental constraints on its final predictions.

On a CVSim-based benchmark, the complete method reduces effect RMSE by **31.87%** and policy-outcome RMSE by **23.07%** relative to a G-Transformer trained on pooled treatment and holdout data. UHC delivers substantial improvements over pooling the same observations, and staged supervision further improves the calibrated model. In the main comparison, ordinary looping and unstaged latent CoT do not significantly improve the calibrated backbone, whereas staged supervision improves both primary metrics. Against unstaged latent CoT with identical architecture, parameter count, and inference computation, its advantage is significant in **each of three training seeds** after multiplicity correction. A separate matched ablation confirms the benefit of the state-then-outcome assignment over uniform and reversed supervision. The complete framework also improves Causal Transformer over its pooled baseline and transfers effect-prediction gains to an unseen policy without retraining or recalibration.

Our contributions combine staged latent computation with experimental trajectory supervision:

- **Staged latent supervision.** We introduce a state-then-outcome supervision schedule for successive latent passes, with final trajectory predictions trained by UHC. This design improves counterfactual accuracy over matched unstaged latent CoT across three training seeds; a separate ablation establishes the benefit of its assignment over uniform and reversed supervision.
- **Universal Holdout Calibration.** We introduce a trajectory-learning framework that turns randomized long-term holdouts into subgroup- and time-specific supervision through state and outcome contrasts and within-arm distributional constraints. This experimental calibration substantially improves counterfactual prediction beyond pooling the same observations.

2 Related Work

Recurrent latent computation and staged supervision. Universal Transformers (Dehghani et al., 2019) introduce recurrent self-attention with shared parameters; Coconut (Hao et al., 2025) propagates continuous hidden states instead of language tokens. Latent Chain-of-Thought Improves Structured-Data Transformers (Dudley & Oymak, 2026) applies latent feedback between Transformer evaluations to tabular and time-series prediction, while Looped World Models (Lu et al., 2026) refine latent states through shared Transformer computation. Intermediate supervision assigns predictive roles to such representations: Deeply-Supervised Nets (Lee et al., 2015) attach auxiliary objectives to hidden layers; LOTUS (Fan et al., 2026) supervises latent positions with explicit reasoning-step targets in a looped architecture; and LCDrive (Tan et al., 2026) interleaves action proposals with latent world representations supervised through future scene rollouts. Our supervision schedule instead follows longitudinal counterfactual estimation: early passes receive state-prediction supervision, later passes receive outcome-prediction supervision, and effects computed from final rollouts are calibrated against randomized subgroup contrasts. Stages communicate through latent representations; observed trajectories provide intermediate targets without individual counterfactual

labels. Our matched ablation tests this task-specific assignment against uniform and reversed supervision, with experimental effect calibration held fixed (Section 6.3).

Sequential counterfactual learning. CRN (Bica et al., 2020) and Causal Transformer (Melnychuk et al., 2022) learn treatment-balanced longitudinal representations, while Mamba-CDSP (Wang et al., 2025) combines state-space modeling with treatment–history decorrelation. G-Net (Li et al., 2021) and G-Transformer (Xiong et al., 2024) model covariate evolution and simulate history-dependent policies through g-computation. Control-based alternatives include SyncTwin (Qian et al., 2021), which constructs synthetic controls from pretreatment representations, and Synthetic Blips (Agarwal et al., 2026), which identifies dynamic effects through latent-factor structure and cross-unit comparisons. We instead use randomized long-term holdouts to constrain learned trajectories with subgroup-level policy contrasts.

Experimental grounding and causal calibration. Experimental Grounding (Kallus et al., 2018) corrects hidden-confounding bias in observational effect models using experimental data. Causal Isotonic Calibration (van der Laan et al., 2023) uses doubly robust targets; Multi-Calibrated Heterogeneous Treatment Effect Estimation (Yuan, 2023) aligns effect predictions with experimental subgroup averages; and Bellman Calibration (van der Laan & Kallus, 2025) recalibrates value predictions against estimated Bellman targets. UHC turns randomized long-term holdouts into subgroup- and time-specific supervision for recursively generated state and outcome trajectories. It combines experimental contrast and within-arm distributional calibration with factual and free-rollout prediction.

Learning under generated histories. Data as Demonstrator (DaD) (Venkatraman et al., 2015) reduces compounding prediction errors by iteratively augmenting training data with model-generated inputs paired with observed next-state targets. UHC adapts supervision under model-generated histories to the treatment–holdout setting, supervising free-running predictions against observed trajectories in both randomized arms and combining this supervision with subgroup contrast and distributional calibration.

3 Problem Setup

Observations and policies. For individuals $i = 1, \dots, n$ at times $t = 0, \dots, T$, B_i denotes pretreatment covariates, $X_{i,t}$ the evolving state, $Y_{i,t}$ the outcome, and $A_{i,t}$ the intervention type and dose. The pre-intervention history is $H_{i,t} = (X_{i,0:t}, Y_{i,0:t}, A_{i,0:t-1})$. Before follow-up, individuals are randomly assigned to the Universal Treatment Group (UTG; $Z_i = 1$) or Universal Control Group (UCG; $Z_i = 0$), following policies π and π_{ref}, respectively. Assignment remains fixed. Dynamic policies select actions, including no-op, using decision-time information and allow time-varying or stochastic decisions. Thus, program assignment—not each treatment decision—is randomized. In our experiments, the reference policy always prescribes no intervention from the target program.

Prediction targets. Potential states and outcomes $X^{\rho}_{i,t}, Y^{\rho}_{i,t}$ under $\rho \in \{\pi, \pi_{\text{ref}}\}$ start from the same observed $(X_{i,0}, Y_{i,0})$. Let U_i denote fixed unobserved individual characteristics and ξ future environmental and policy randomness. We target expected outcomes and individual-specific policy effects:

$$\mu^{\rho}_{i,t} = \mathbb{E}_{\xi}\left[Y^{\rho}_{i,t}(B_i, U_i, \xi)\right], \qquad \tau_{i,t} = \mu^{\pi}_{i,t} - \mu^{\pi_{\text{ref}}}_{i,t}.$$

Expectation averages over ξ, holding B_i, U_i and observed initial conditions fixed. The learner predicts these quantities from observed information; subsequent interventions depend on each rollout's generated history.

Training information. Data $\mathcal{D} = \{B_i, Z_i, X_{i,0:T}, Y_{i,0:T}, A_{i,0:T-1}\}_{i=1}^{n}$ contain only each individual's assigned-program trajectory. Neither U_i, unrealized trajectories, nor individual effect labels are available. Individual-specific prediction targets therefore differ from available training supervision. Simulator-derived counterfactual outcomes and expected effects are reserved for held-out evaluation, never fitting or model selection.

Identification. We assume randomized program assignment, positive probability of each arm, consistency of observed and assigned-policy potential trajectories, and no interference between individuals. A common fixed pretreatment rule $G_i = g(B_i)$ defines subgroups. Randomized comparisons identify $\tau_{g,t} = \mathbb{E}[\tau_{i,t} \mid G_i = g]$, the subgroup conditional average policy effect, without observ-

ing every individual characteristic affecting treatment response. This identifies the implemented program-versus-reference contrast, not individual effects or arbitrary untested policies. Section 5 derives observed subgroup state and outcome contrasts used as anchors.

Recursive learning. Our G-Transformer implementation learns $p_\theta(X_{i,t+1}, Y_{i,t+1} \mid H_{i,t}, A_{i,t})$. Starting from common observed initial conditions, each policy rollout alternates intervention selection and state–outcome prediction. Averaging simulated trajectories yields predicted expected outcomes; their difference yields the policy-effect estimate. Staged Latent Chain-of-Thought assigns state- and outcome-specific objectives to successive latent passes (Section 4). UHC trains the final trajectory predictions through factual learning, free-rollout supervision, and subgroup calibration (Section 5).

4 Staged Latent Chain-of-Thought

4.1 Structuring latent computation around the prediction task

Staged Latent Chain-of-Thought combines intermediate supervision from observed states and outcomes with UHC, which trains final trajectories against individual observations and randomized group comparisons.

The supervision schedule is motivated by three connected tasks: predicting how interventions change states, predicting outcomes under those evolving states, and estimating effects by comparing policy and reference trajectories. Errors in state and outcome predictions therefore enter the final policy-effect estimate. We assign next-state supervision to an early latent pass and next-outcome supervision to the subsequent pass, while retaining joint state–outcome prediction and trajectory-level calibration at the final pass. Here, stages refer to the assignment of prediction-specific supervision to successive latent passes within each transition.

4.2 Latent feedback within each transition

We use the appended-feedback mechanism of latent chain-of-thought for structured-data Transformers (Dudley & Oymak, 2026). An initial pass followed by two feedback updates gives recurrence depth $R = 2$ and three passes, $r = 0, 1, 2$. Encoder parameters are shared across passes; each pass receives the original embedded history and all preceding feedback tokens. Feedback consists of transformed hidden representations, not decoded state or outcome predictions. State supervision shapes the representation from which the next latent pass receives information. Latent recurrence refines a single transition without advancing physical time; temporal rollout instead constructs the next history from the final transition prediction. The three-pass computation is repeated at each rollout step. Appendix D specifies feedback construction and readout.

4.3 Intermediate prediction objectives

Pass $r = 0$ receives the state loss $\mathcal{L}^{(0)}_{X,\text{factual}}$, combining squared-error regression for continuous states with cross-entropy for categorical states. Pass $r = 1$ receives the outcome loss $\mathcal{L}^{(1)}_{Y,\text{factual}}$, using normalized squared error. Both objectives are evaluated only on observed history–action prefixes, using the same next-step targets as the final factual loss. They supervise different representations for one transition, not separate intermediate counterfactual rollouts. Appendix D gives their sample-average definitions, normalization, and shared-head implementation.

4.4 Final transitions, effect calibration, and joint training

The final pass, $r = 2$, jointly predicts states and outcomes and supplies all predictions used by UHC. At evaluation, this final-pass model generates dynamic-policy and reference-policy trajectories, with subsequent dynamic-policy actions selected from the generated history. Policy effects are computed from the difference between the two predicted outcome trajectories, without an independent effect head. Section 5 defines UHC's trajectory-learning objective and calibration targets.

The complete training objective is

$$\boxed{\mathcal{L}_{\text{Stage}} = \mathcal{L}^{(2)}_{\text{UHC}} + \lambda_X \mathcal{L}^{(0)}_{X,\text{factual}} + \lambda_Y \mathcal{L}^{(1)}_{Y,\text{factual}}}.$$

Here, $\mathcal{L}^{(2)}_{\text{UHC}}$ applies the full UHC objective to final-pass predictions. Staged supervision adds state- and outcome-specific objectives at the intermediate passes. Gradients from the final objective propagate through latent feedback, jointly training the passes and their predictions. Appendix D provides auxiliary coefficients and implementation details.

Staged and unstaged Latent CoT use the same architecture, parameter count, and inference computation. Their only difference is intermediate supervision during training, isolating its contribution from that of additional recurrent computation. Complementary ablations keep the intermediate targets and total loss weights fixed while varying their assignment across passes (Appendix H.3).

5 Universal Holdout Calibration

UHC transforms randomized long-term holdouts into subgroup- and time-specific supervision for state and outcome trajectories. It trains a shared trajectory model through experimental mean contrasts, within-arm distributional constraints, and individual prediction losses.

5.1 Joint trajectory-learning objective

UHC jointly optimizes factual prediction, free-rollout supervision in both randomized arms, and experimental calibration:

$$\boxed{\mathcal{L}_{\text{UHC}} = \mathcal{L}_{\text{factual}} + \lambda_0 \mathcal{L}^{\text{UCG}}_{\text{free}} + \lambda_1 \mathcal{L}^{\text{UTG}}_{\text{free}} + \lambda_\Delta \mathcal{L}_{\text{contrast}} + \lambda_d \mathcal{L}_{\text{distribution}}}.$$

Factual learning predicts next states and outcomes from observed histories. Free-rollout learning recursively feeds predictions back into the model and compares them with each individual's observed trajectory. UCG rollouts follow no-op actions; UTG rollouts replay observed actions, with losses normalized separately by arm. This applies supervision under model-generated histories (Venkatraman et al., 2015) to both randomized arms.

All terms update the same trajectory model. Subgroup constraints apply during training; inference requires no comparison cohort or post-hoc correction. Staged Latent CoT applies UHC to its final-pass predictions. Appendix C provides loss normalizations, weights, and sampling details.

5.2 Causal grounding of subgroup contrasts

Let $V_{i,t} = (X_{i,t}^\top, Y_{i,t})^\top$ collect the state and outcome, and let $G_{i,t} = g_t(B_i)$ denote a time-specific subgroup defined by a fixed pretreatment rule common to both arms. Under the assumptions in Section 3, randomized assignment and consistency give

$$\begin{aligned}\Delta_{g,t} &= \mathbb{E}\big[V^{\pi}_{i,t} - V^{\pi_{\text{ref}}}_{i,t} \mid G_{i,t} = g\big] \\ &= \mathbb{E}[V_{i,t} \mid Z_i = 1, G_{i,t} = g] - \mathbb{E}[V_{i,t} \mid Z_i = 0, G_{i,t} = g].\end{aligned}$$

The outcome component is the subgroup CATE; the state components give average policy effects on evolving states. These identified contrasts provide supervision across the trajectory without reconstructing unobserved individual characteristics.

5.3 Constructing and selecting calibration cells

Calibration cells are subgroup–time pairs that separate expected intervention patterns and initial conditions. At each prediction time, we summarize the preceding observed UTG interventions into treatment-profile labels. We train a predictor of these labels using pretreatment covariates and apply it to both arms to define common baseline regions. Within each retained region, refinement by X_0 separates initial states. The resulting routing rules are frozen, and subgroup assignment in both arms uses only pretreatment information.

At construction, we retain cells with sufficient support in both arms, acceptable pretreatment balance and outcome-contrast precision, and at least one non-no-op intervention in a UTG member's prefix. These checks select calibration constraints; all individuals remain in factual training. Appendix B specifies the construction folds, profile filters, refinement, and acceptance criteria.

5.4 Trajectory-wide contrast and distributional targets

Mean-contrast targets. For a training mini-batch $\mathcal{B}$, let $\mathcal{I}^{\mathcal{B}}_{a,g,t} = \{i \in \mathcal{B} : Z_i = a,\ G_{i,t} = g\}$ and $n^{\mathcal{B}}_{a,g,t} = |\mathcal{I}^{\mathcal{B}}_{a,g,t}|$ denote cell membership and arm counts. When both arms are represented, the observed anchor is

$$\widehat{\Delta}^{\mathcal{B}}_{g,t} = \frac{1}{n^{\mathcal{B}}_{1,g,t}} \sum_{i \in \mathcal{I}^{\mathcal{B}}_{1,g,t}} V_{i,t} - \frac{1}{n^{\mathcal{B}}_{0,g,t}} \sum_{i \in \mathcal{I}^{\mathcal{B}}_{0,g,t}} V_{i,t}.$$

For each UTG individual, two rollouts start from the same observed initial conditions: $\widetilde{V}^{\text{rep}}_{i,t}$ replays observed interventions, and $\widetilde{V}^{0}_{i,t}$ follows the no-op reference. Their subgroup-average contrast is

$$\widehat{d}^{\mathcal{B}}_{\theta,g,t} = \frac{1}{n^{\mathcal{B}}_{1,g,t}} \sum_{i \in \mathcal{I}^{\mathcal{B}}_{1,g,t}} \left(\widetilde{V}^{\text{rep}}_{i,t} - \widetilde{V}^{0}_{i,t} \right).$$

The identified policy contrast supervises the model's paired replay contrast through a subgroup-level surrogate loss. Evaluation selects actions from generated histories; randomization does not establish replay–policy equivalence (Appendix A.7).

The contrast objective is

$$\mathcal{L}_{\text{contrast}} = \sum_{(g,t) \in \mathcal{C}_{\mathcal{B}}} w_{g,t} \left[\left(\widehat{d}^{Y}_{\theta,g,t} - \widehat{\Delta}^{Y}_{g,t} \right)^2 + \frac{1}{d_X} \left\| \widehat{d}^{X}_{\theta,g,t} - \widehat{\Delta}^{X}_{g,t} \right\|_2^2 \right],$$

where $\mathcal{C}_{\mathcal{B}}$ contains retained cells represented in both arms of $\mathcal{B}$, d_X is the state dimension, and $w_{g,t}$ are normalized weights accounting for the number of retained cells at each time. Continuous targets and predictions share training-fitted normalization; categorical indicators are compared with predicted probabilities.

Within-arm distributional targets. Mean-contrast calibration aligns subgroup-average rollout contrasts with observed between-arm differences. Within-arm distributional targets additionally constrain the variability, tails, and contemporaneous dependence of generated states and outcomes. We match variances, quantiles, tail statistics, and contemporaneous state–state and state–outcome correlations of continuous variables within each subgroup, arm, and time. These targets are fixed from observed reference-fold data, whereas mean-contrast anchors are recomputed in mini-batches from the full source-training cohort. Stochastic rollouts replay observed interventions and match their within-arm statistics to the fixed targets. Appendix C specifies support requirements, sampling, and Monte Carlo aggregation.

6 Experiments

We first evaluate whether staged latent supervision improves sequential counterfactual prediction beyond UHC alone, ordinary looping, and unstaged latent chain-of-thought. We then assess the gains from the integrated UHC objective over pooling the same observations, the transfer of the complete framework to Causal Transformer, and its generalization to an unseen policy.

6.1 Experimental Setup

Benchmark and data-generating process. We construct a CVSim-based benchmark (Heldt et al., 2010) for predicting trajectories under a dynamic intervention policy and a no-op reference policy. The outcome Y is mean arterial pressure (MAP), and the state X contains the remaining 17 measurements, comprising 16 continuous variables and one binary variable. At each decision time, the intervention policy selects no-op, fluid administration, or vasopressor administration, including

the corresponding dose. Interventions affect subsequent physiology, which in turn determines later intervention decisions. We evaluate trajectories over $T = 16$ prediction steps. We strengthen treatment effects, increase treatment coverage, and introduce patient-specific latent response variation unavailable to the models, creating substantive policy–no-op contrasts and unobserved response heterogeneity. Appendix E specifies the simulator configuration.

Cohorts and available information. Both potential paths undergo quality-control screening before random assignment to UCG or UTG. Training receives only each patient's assigned-arm trajectory. The source-training cohort contains 90,000 patients, equally divided between arms; validation contains 1,000 patients. All models use the same 5,997 independent test patients after an additional reference-trajectory screen that does not use model predictions or errors. Appendix E documents both screening stages. Simulator-derived expected outcomes and policy effects are reserved for evaluation, with reference-trajectory construction detailed in Appendix G.

Comparison models and task adaptations. Pooled GT, Pooled CT, and Pooled Mamba-CDSP train on the combined UCG and UTG data, treating UCG trajectories as factual no-op observations with zero doses, without subgroup calibration. Calibrated GT (full UHC) augments factual prediction with individual-level free-rollout supervision in both randomized arms, subgroup contrast calibration, and within-arm distributional calibration (Section 5). Its three extensions add block-wise looping with four repetitions, ordinary latent chain-of-thought, or staged latent supervision as defined in Section 4. These form the seven-model main comparison. Separately, we compare Pooled CT with Calibrated CT + Staged Latent CoT, using the same calibration configuration as Calibrated GT. GT already predicts states and outcomes jointly. For CT and Mamba-CDSP, we add state-prediction heads while retaining their backbones and treatment-balancing or CDSP objectives. Both CT variants receive the same task adaptation. All models represent intervention type and dose and evaluate the dynamic policy on generated histories, without future observed states, outcomes, or actions. Appendix D gives architectural details; Appendix G describes the rollouts.

As an additional seed-2032 ablation, Anchor GT augments factual prediction with subgroup state and outcome contrasts and within-arm distributional calibration, omitting individual-level free-rollout losses in both arms (Appendix H.2). We also compare uniform and reversed intermediate supervision against the proposed schedule at seed 2032 (Appendix H.3).

Training and model selection. All models share the source observations, observed-variable information, validation cohort, and test references. Training is capped at 80 epochs, with batch size 512 for GT and CT and 128 for Mamba-CDSP. Checkpoints are selected by factual prediction loss on observed validation data, never by counterfactual test performance. The two CT variants also share optimizer hyperparameters and the selection criterion. Appendix F specifies backbone-specific optimization settings.

Evaluation metrics and uncertainty. Our primary metrics, Effect RMSE and Policy Y RMSE, measure individual-specific expected policy effects and expected dynamic-policy outcomes across all test patients and prediction times. We use paired patient-bootstrap intervals conditional on fixed checkpoints. Model-specific intervals are pointwise 95%. Bonferroni correction covers the six alternatives in the seven-model main comparison, separately within each primary metric. Anchor GT, supervision-schedule ablations, and CT transfer are evaluated separately from this family. Appendix G specifies reference estimation, supplementary metrics, and the bootstrap procedure; Appendix H reports detailed results.

6.2 Overall Performance

Staged Latent CoT with UHC achieves the lowest Effect RMSE and Policy Y RMSE among the seven models in the main comparison (Table 1), reducing these errors by **31.87%** and **23.07%** relative to Pooled GT. Its advantage over all six alternatives is significant on both primary metrics after Bonferroni correction (Table H.4). Supplementary metrics and terminal Effect RMSE appear in Tables H.1 and H.5.

Without retraining or recalibration, the complete GT method also generalizes to the prespecified unseen MAP-target-80 policy, reducing all-time and terminal Effect RMSE by **13.39%** and **33.36%** relative to Pooled GT; the paired 95% confidence intervals for both differences exclude zero (Appendix J).

Table 1: Primary prediction metrics on the 5,997-patient test cohort. The main comparison comprises the seven models without †; Anchor GT† is an additional ablation evaluated separately. Entries are point estimates with pointwise 95% paired patient-bootstrap confidence intervals. Lower is better; bold denotes the lowest point estimate among the displayed models.

Model	Effect RMSE, all times ↓	Policy Y RMSE ↓
Pooled GT	14.5826 [14.2731, 14.8974]	12.8461 [12.5666, 13.1323]
Anchor GT†	10.3922 [10.0689, 10.7200]	10.4084 [10.0868, 10.7350]
Calibrated GT (full UHC)	10.1987 [9.9295, 10.4734]	10.0769 [9.8054, 10.3530]
Calibrated GT + Loop	10.1611 [9.8782, 10.4478]	10.0913 [9.8104, 10.3768]
Calibrated GT + Latent CoT	10.2410 [9.9623, 10.5255]	10.1399 [9.8612, 10.4245]
Calibrated GT + Staged Latent CoT	**9.9358** [9.6477, 10.2283]	**9.8819** [9.5955, 10.1733]
Pooled CT	11.2789 [10.9063, 11.6556]	11.9516 [11.5765, 12.3276]
Pooled Mamba-CDSP	10.9792 [10.6392, 11.3249]	10.9560 [10.6160, 11.3032]

6.3 Effects of Staged Latent Supervision and Integrated UHC

Staged latent supervision. With the UHC objective held fixed, staged latent supervision significantly improves both primary metrics over Calibrated GT, ordinary looping, and unstaged latent chain-of-thought in the main comparison. Relative to Calibrated GT, Effect RMSE decreases by **2.58%** and Policy Y RMSE by **1.94%**. The paired differences are -0.2629 [-0.3735, -0.1469] and -0.1950 [-0.3039, -0.0820], respectively, with Bonferroni-adjusted 95% intervals (Table H.4). Ordinary looping and unstaged latent CoT do not significantly improve either primary metric over Calibrated GT (Table H.3).

The matched comparison evaluates the added value of the proposed intermediate state and outcome supervision: the UHC objective, architecture, parameter count, and inference computation are unchanged. Adding these objectives significantly improves both primary metrics over unstaged latent CoT in **all three training seeds** after Bonferroni correction within each seed and metric. Across-seed mean Effect RMSE and Policy Y RMSE decrease by **2.88%** and **2.50%**, respectively (Appendix I). The looped model's lower no-op RMSE and one-step state NRMSE do not translate into better policy-effect prediction (Table H.1).

Supervision assignment and order. At training seed 2032, state-then-outcome supervision reduces Effect RMSE and Policy Y RMSE by **2.91%** and **2.27%** relative to uniform supervision of both targets at both intermediate passes, and by **2.82%** and **2.29%** relative to the reversed schedule. All four differences remain significant after correction across 36 comparisons. With targets, total loss weights, initialization, architecture, and inference computation held fixed, these comparisons establish the benefit of the proposed assignment and order (Appendix H.3).

Integrated UHC. Relative to Pooled GT, Anchor GT reduces Effect RMSE by **28.74%** and Policy Y RMSE by **18.98%** without individual-level free-rollout supervision. These gains demonstrate the effectiveness of combined subgroup contrast and distributional calibration beyond pooled factual training. Calibrated GT (full UHC) adds free-rollout losses in both arms and further lowers these point estimates by **1.86%** and **3.18%**. Both primary RMSEs decrease progressively across Pooled GT, Anchor GT, Calibrated GT, and the staged model in this seed-2032 comparison (Table 1; Appendix H.2). Under dynamic-policy evaluation, Calibrated GT reduces subgroup mean-effect RMSE from **10.4021 to 2.3593 mmHg** relative to Pooled GT, with lower absolute error in all 26 training-fixed cells (Appendix H.1). Paired primary-metric differences appear in Table H.3.

6.4 Transfer to Causal Transformer

Applying the combined framework to CT reduces Effect RMSE by **13.04%** and Policy Y RMSE by **18.28%** relative to Pooled CT (Table 2). Both improvements are significant in the paired comparison (Table H.3). The matched data, optimization, and checkpoint-selection protocol is specified in Section 6.1. This experiment evaluates the joint transfer of calibration and staged supervision to another backbone. Supplementary and terminal metrics appear in Tables H.2 and H.6.

Table 2: Transfer of the combined framework to Causal Transformer. Entries are point estimates with pointwise 95% paired patient-bootstrap confidence intervals on the same test cohort. Lower is better; bold denotes the lower CT point estimate.

Model	**Effect RMSE, all times** ↓	**Policy Y RMSE** ↓
Pooled CT	11.2789 [10.9063, 11.6556]	11.9516 [11.5765, 12.3276]
Calibrated CT + Staged Latent CoT	**9.8077** [9.4954, 10.1242]	**9.7674** [9.4576, 10.0830]

7 Discussion and Conclusion

Staged Latent CoT with Universal Holdout Calibration (UHC) combines two complementary contributions to sequential counterfactual prediction: task-aligned supervision of intermediate latent computation and experimental calibration of generated trajectories. State-then-outcome supervision structures computation within each transition, while UHC turns long-term randomized holdouts into subgroup- and time-specific constraints on state and outcome trajectories. Training requires no individual counterfactual labels, and inference requires no comparison cohort.

With UHC, architecture, parameter count, and inference computation fixed, staged supervision significantly improves Effect RMSE and Policy Y RMSE over unstaged Latent CoT in each of three training seeds after multiplicity correction, reducing across-seed mean RMSEs by 2.88% and 2.50%. The staged model also improves on non-recurrent Calibrated GT in every seed by point estimate, significantly in two, whereas neither ordinary looping nor unstaged Latent CoT significantly improves it. A separate seed-2032 ablation holds intermediate targets and total loss weights fixed: state-then-outcome supervision significantly outperforms uniform and reversed assignments on both metrics after correction across 36 comparisons. Both alternatives remain close to the unstaged level, indicating that the gain depends on how the auxiliary targets are assigned across passes, not simply on their inclusion. The alternative schedules improve one-step state accuracy yet worsen counterfactual prediction, showing that better local prediction does not ensure more accurate policy rollouts.

UHC combines between-arm mean contrasts and within-arm distributional constraints with factual and free-rollout learning. Even without individual-level free-rollout losses, joint contrast and distributional calibration reduces Effect RMSE by 28.74% and Policy Y RMSE by 18.98% relative to Pooled GT, amounting to 96% and 88% of the full-UHC reductions. The complete framework significantly outperforms all six main-comparison alternatives on both primary metrics after correction, including the strongest pooled baseline, Mamba-CDSP, by 9.50% and 9.80%. It also improves Causal Transformer over its pooled baseline by 13.04% and 18.28%, and transfers effect-prediction gains to one unseen policy without retraining or recalibration.

This evidence spans two backbones and one unseen policy within a single CVSim-based benchmark at a fixed recurrence depth. Randomization identifies population targets that replay-based training uses as surrogate regularization; under direct dynamic-policy evaluation, UHC still reduces subgroup mean-effect RMSE on independent test patients from 10.40 to 2.36 mmHg, with lower error in all 26 calibration cells. Deeper recurrences and additional longitudinal tasks are direct extensions of this design. Together, these contributions open a research direction: jointly designing experimental constraints on generated trajectories and supervision of the latent computation that produces them.

AI Use Statement

We used GPT-6 Astra, Codex, and Claude Opus 5.5 for coding assistance, literature research, translation, and drafting and revising English text. We also used these tools to aggregate and summarize experimental results, prepare tables and research materials, and obtain simulated reviewer feedback on the manuscript, including its presentation and interpretation of results. The author takes full responsibility for the final content, methods, results, and claims of this work.

REFERENCES

Anish Agarwal, Sukjin Han, Dwaipayan Saha, Vasilis Syrgkanis, and Haeyeon Yoon. Synthetic blips: Generalizing synthetic controls for dynamic treatment effects. arXiv preprint arXiv:2210.11003v3, 2026. URL https://arxiv.org/abs/2210.11003v3. Revised September 11, 2026; first submitted in 2022.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJg866NFvB.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Carson Dudley and Samet Oymak. Latent chain-of-thought improves structured-data transformers. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=L0Mj9fY94c.

Ying Fan, Anej Svete, and Kangwook Lee. Bridging the gap between latent and explicit reasoning with looped transformers. arXiv preprint arXiv:2606.31779, 2026. URL https://arxiv.org/abs/2606.31779.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E. Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Thomas Heldt, Ramakrishna Mukkamala, George B. Moody, and Roger G. Mark. CVSim: An open-source cardiovascular simulator for teaching and research. *The Open Pacing, Electrophysiology & Therapy Journal*, 3:45–54, 2010. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3178445/.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10888–10897, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/566f0ea4f6c2e947f36795c8f58ba901-Abstract.html.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 562–570. PMLR, 2015. URL https://proceedings.mlr.press/v38/lee15a.html.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M. Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, and Li-wei Lehman. G-Net: A recurrent network approach to G-Computation for counterfactual prediction under a dynamic treatment regime. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 282–299. PMLR, 2021. URL https://proceedings.mlr.press/v158/li21a.html.

Hongyuan Adam Lu, Z. L. Victor Wei, Qun Zhang, Jinrui Zeng, Bowen Cao, Lingwei Meng, Mocheng Li, Zezhong Wang, Haonan Yin, Naifu Xue, Minyu Chen, Cenyuan Zhang, Zefan Zhang, Hao Wei, Jiawei Zhou, Haoran Xu, Hao Yang, Ronglai Zuo, Tongda Xu, Yonghao Li, Jian Chen, Hebin Wang, Zeyu Gao, Yang Li, Wei Zhao, Qimin Zhong, Siqi Liu, Yumeng Zhang, Leyan Cui, Zhangyu Wang, and Wai Lam. Looped world models. arXiv preprint arXiv:2606.18208, 2026. URL https://arxiv.org/abs/2606.18208. Technical report.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal Transformer for estimating counterfactual outcomes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15293–15329. PMLR, 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

Zhaozhi Qian, Yao Zhang, Ioana Bica, Angela Wood, and Mihaela van der Schaar. SyncTwin: Treatment effect estimation with longitudinal outcomes. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3178–3190, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/19485224d128528da1602ca47383f078-Abstract.html.

Kevin Shectman. Using global control groups to measure the impact of customer engagement. Braze, May 2021. URL https://www.braze.com/resources/articles/using-global-control-groups-to-measure-the-impact-of-customer-engagement. Accessed September 20, 2026.

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krähenbühl, Marco Pavone, and Boris Ivanovic. Latent chain-of-thought world modeling for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39724–39733, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Tan_Latent_Chain-of-Thought_World_Modeling_for_End-to-End_Autonomous_Driving_CVPR_2026_paper.html.

The SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *The New England Journal of Medicine*, 373(22):2103–2116, 2015. doi: 10.1056/NEJMoa1511939. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4689591/.

Lars van der Laan and Nathan Kallus. Bellman calibration for V-learning in offline reinforcement learning. arXiv preprint arXiv:2512.23694, 2025. URL https://arxiv.org/abs/2512.23694v2. Revised May 7, 2026 (version 2).

Lars van der Laan, Ernesto Ulloa-Perez, Marco Carone, and Alex Luedtke. Causal isotonic calibration for heterogeneous treatment effects. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34831–34854. PMLR, 2023. URL https://proceedings.mlr.press/v202/van-der-laan23a.html.

Arun Venkatraman, Martial Hebert, and J. Andrew Bagnell. Improving multi-step prediction of learned time series models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 3024–3030, 2015. doi: 10.1609/aaai.v29i1.9590. URL https://ojs.aaai.org/index.php/AAAI/article/view/9590.

Haotian Wang, Haoxuan Li, Hao Zou, Haoang Chi, Long Lan, Wanrong Huang, and Wenjing Yang. Effective and efficient time-varying counterfactual prediction with state-space models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/26f16a11f4dbba3eef876b571d9f3200-Abstract-Conference.html.

Hong Xiong, Feng Wu, Leon Deng, Megan Su, Zach Shahn, and Li-wei H. Lehman. G-Transformer: Counterfactual outcome prediction under dynamic and time-varying treatment regimes. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*, pp. 1–28. PMLR, 2024. URL https://proceedings.mlr.press/v252/xiong24a.html.

Yuan Yuan. Multi-calibrated heterogeneous treatment effect estimation. Working paper, version dated December 28, 2023, 2023. URL https://www.yuan-yy.com/HTE_cali_202312.pdf.

Appendix

Experimental Rationale and Interpretation

Additional value of staged supervision. With the UHC objective, architecture, parameter count, and inference computation held fixed, adding the proposed intermediate state and outcome supervision significantly improves both primary metrics over unstaged Latent CoT in all three training seeds after Bonferroni correction within each seed and metric (Appendix I). In the main comparison, ordinary looping and unstaged Latent CoT do not significantly improve Calibrated GT, whereas staged supervision significantly outperforms all three alternatives (Tables H.3 and H.4). Staged supervision also achieves lower point estimates than Calibrated GT on both primary metrics in every seed (Table I.1); these differences are significant after correction in seeds 2031 and 2032, but not in seed 2033 (Table I.2).

Rationale for the supervision schedule. The state-then-outcome schedule is motivated by the structure of sequential counterfactual prediction: evolving states shape subsequent outcomes and policy decisions, and predicted trajectories determine policy-effect estimates. We therefore assign state and outcome objectives to successive latent passes while retaining joint prediction at the final pass. Beyond the three-seed comparison against unstaged Latent CoT, a matched seed-2032 ablation holds the intermediate targets and total loss weights fixed and varies their assignment across passes. The proposed schedule significantly outperforms uniform and reversed supervision on both primary metrics after correction across 36 comparisons (Appendix H.3). This directly supports the state-then-outcome assignment, in addition to the benefit of intermediate supervision as a whole.

Role of the integrated UHC objective. UHC provides the shared trajectory-learning foundation for staged and unstaged Latent CoT. The Anchor GT ablation shows that combined contrast and distributional calibration improves both primary metrics over pooled factual training without individual-level free-rollout supervision. Full UHC achieves lower primary RMSE point estimates than Anchor GT in the seed-2032 comparison (Appendix H.2). With this complete objective held fixed, the matched staged-versus-unstaged comparison establishes the additional benefit of the proposed intermediate supervision across three training seeds.

A Causal Basis of Holdout Calibration

Universal Holdout Calibration uses randomized program comparisons to construct supervision for state and outcome trajectories. This appendix establishes the causal basis of these targets. We distinguish three objects: the population policy contrasts identified by randomization, the empirical anchors estimated from observed trajectories, and the model predictions trained against those anchors. This distinction makes explicit what the experimental comparison identifies and how that information enters trajectory learning.

A.1 Randomized program assignment and the target contrast

Randomization occurs at the level of program assignment. Let $Z_i = 1$ indicate assignment to the intervention program and $Z_i = 0$ assignment to its long-term holdout. Individuals assigned to the program follow the dynamic policy π, while holdout individuals follow the reference policy π_{ref}. In our experiments, the reference policy prescribes no intervention from the target program.

Program assignment is distinct from treatment receipt at an individual time step. Within UTG, the policy selects interventions using the evolving history and can prescribe no intervention. The experimental contrast therefore measures the effect of following the intervention program rather than the reference policy, averaging over the treatment sequences generated by each policy.

For each policy $\rho \in \{\pi, \pi_{\text{ref}}\}$, define the potential state–outcome vector

$$V_{i,t}^{\rho} = \begin{pmatrix} X_{i,t}^{\rho} \\ Y_{i,t}^{\rho} \end{pmatrix},$$

and let $V_{i,t}$ denote its observed counterpart. At prediction time t, subgroup membership is defined by

$$G_{i,t} = g_t(B_i),$$

where B_i contains pretreatment covariates and g_t is a fixed rule applied identically to both arms. Different prediction times can use different rules, but each rule uses only pretreatment information.

For a subgroup with positive population probability, the target contrast is

$$\Delta_{g,t} = \mathbb{E}\left[V_{i,t}^{\pi} - V_{i,t}^{\pi_{\text{ref}}} \mid G_{i,t} = g\right].$$

Its outcome component is the subgroup conditional average policy effect. Its state components measure the corresponding average effects on the evolving state variables. These targets concern the entire assigned policy, not treatment receipt or a particular realized treatment sequence within the subgroup.

A.2 Identification assumptions

We use the following assumptions.

1. **Randomized program assignment.** Assignment is independent of pretreatment characteristics and potential trajectories:

$$Z_i \perp\!\!\!\perp \left(B_i,\, U_i,\, V_{i,0:T}^{\pi},\, V_{i,0:T}^{\pi_{\text{ref}}}\right),$$

   where U_i denotes fixed, unobserved individual characteristics. Potential trajectories include the subsequent environmental and policy randomness relevant to each policy.

2. **Positive assignment probabilities.** Both arms have positive probability within each subgroup used for identification:

$$0 < \Pr(Z_i = 1 \mid G_{i,t} = g) < 1.$$

   Under the unconditional randomization above, this follows from positive assignment probabilities for both arms and positive subgroup probability.

3. **Consistency.** The observed trajectory equals the potential trajectory under the assigned policy:

$$V_{i,t} = \begin{cases} V_{i,t}^{\pi}, & Z_i = 1, \\ V_{i,t}^{\pi_{\text{ref}}}, & Z_i = 0. \end{cases}$$

4. **No interference between individuals.** An individual's potential trajectory under a policy does not depend on the program assignments of other individuals.

The fixed, common subgroup rule is essential to the argument. Conditioning on $g_t(B_i)$ restricts the comparison using information determined before intervention. It does not select individuals according to their realized post-intervention states, outcomes, or treatment receipt.

Comparability within a subgroup does not mean that its members have identical characteristics or responses. It means that assignment remains independent of their potential trajectories after conditioning on subgroup membership.

A.3 Identification of subgroup policy effects

Proposition 1: Subgroup policy-contrast identification. Under the assumptions above, for every fixed pretreatment-defined subgroup with positive probability,

$$\boxed{\Delta_{g,t} = \mathbb{E}[V_{i,t} \mid Z_i = 1, G_{i,t} = g] - \mathbb{E}[V_{i,t} \mid Z_i = 0, G_{i,t} = g].}$$

Proof. Because $G_{i,t}$ is a fixed function of B_i, randomized assignment implies

$$Z_i \perp\!\!\!\perp \left(V_{i,t}^{\pi}, V_{i,t}^{\pi_{\text{ref}}}\right) \mid G_{i,t}.$$

For the intervention arm, consistency and conditional independence give

$$\begin{aligned}\mathbb{E}[V_{i,t} \mid Z_i = 1, G_{i,t} = g] &= \mathbb{E}[V_{i,t}^{\pi} \mid Z_i = 1, G_{i,t} = g] \\ &= \mathbb{E}[V_{i,t}^{\pi} \mid G_{i,t} = g].\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}[V_{i,t} \mid Z_i = 0, G_{i,t} = g] &= \mathbb{E}[V_{i,t}^{\pi_{\text{ref}}} \mid Z_i = 0, G_{i,t} = g] \\ &= \mathbb{E}[V_{i,t}^{\pi_{\text{ref}}} \mid G_{i,t} = g].\end{aligned}$$

Subtracting the two expressions establishes the result. □

For the outcome component, the proposition connects directly to the individual-specific expected effects defined in Section 3. Recall that

$$\mu_{i,t}^{\rho} = \mathbb{E}_{\xi}\left[Y_{i,t}^{\rho}(B_i, U_i, \xi)\right], \qquad \tau_{i,t} = \mu_{i,t}^{\pi} - \mu_{i,t}^{\pi_{\text{ref}}},$$

where the expectation averages future randomness while holding individual characteristics fixed. By the law of iterated expectations,

$$\Delta_{g,t}^{Y} = \mathbb{E}\left[\mu_{i,t}^{\pi} - \mu_{i,t}^{\pi_{\text{ref}}} \mid G_{i,t} = g\right] = \mathbb{E}[\tau_{i,t} \mid G_{i,t} = g].$$

Thus, the observed subgroup contrast identifies the average of the individual-specific expected policy effects within that subgroup.

The same argument applies at every prediction time. It therefore supplies experimental targets along the trajectory, including intermediate state changes and outcomes, rather than only at the terminal horizon.

A.4 Why unobserved individual characteristics do not invalidate the contrast

The identification argument does not require the learner to observe every characteristic influencing treatment response. Randomized assignment also implies

$$\mathcal{L}(U_i \mid Z_i = 1, G_{i,t} = g) = \mathcal{L}(U_i \mid Z_i = 0, G_{i,t} = g) = \mathcal{L}(U_i \mid G_{i,t} = g),$$

where $\mathcal{L}$ denotes a probability distribution.

Unobserved characteristics can influence state evolution, outcomes, and the magnitude of intervention responses. Their contributions remain part of the potential trajectories and hence part of the policy-effect estimand. Randomization ensures that the intervention and holdout arms average over the same distribution of these characteristics within the subgroup. The causal comparison therefore does not depend on reconstructing U_i from observed inputs.

This argument concerns program assignment, not the assignment of each treatment within the program. Treatment decisions under the dynamic policy can be associated with evolving states and response determinants. Identifying the subgroup contrast between the two randomized policies does not require those individual decisions to be randomized, nor does it require sequential exchangeability given the learner's observed history.

The resulting supervision is a subgroup-average policy contrast. It does not identify every individual's effect or the effects of arbitrary policies not represented by the experiment. Individual effect prediction remains a modeling task, constrained by experimental averages and learned trajectory structure. The causal guarantee applies to the population targets supplied by the randomized comparison.

A.5 Identification of within-policy distributional targets

Randomization identifies more than subgroup means. For $\rho_1 = \pi$ and $\rho_0 = \pi_{\text{ref}}$, the same assumptions imply

$$\mathcal{L}(V_{i,t} \mid Z_i = z, G_{i,t} = g) = \mathcal{L}(V_{i,t}^{\rho_z} \mid G_{i,t} = g), \qquad z \in \{0, 1\}.$$

Equivalently, for any measurable function h with a finite expectation,

$$\mathbb{E}[h(V_{i,t}) \mid Z_i = z, G_{i,t} = g] = \mathbb{E}[h(V_{i,t}^{\rho_z}) \mid G_{i,t} = g].$$

Choosing h as a coordinate, squared coordinate, or product of coordinates yields first moments, second moments, and cross-moments. Indicator functions identify probabilities and distribution functions. These relationships provide the basis for calibrating within-arm variances, quantiles, tail behavior, and contemporaneous X–X and X–Y correlations.

The mean-contrast and distributional targets have distinct roles. Mean contrasts constrain how the intervention program changes subgroup-average states and outcomes relative to the holdout. Within-arm targets constrain the distributional structure of the trajectories generated under each policy. Both derive from the randomized comparison, but they impose different requirements on the learned predictions.

Here, a within-policy joint distribution includes relationships among state and outcome variables under the same policy. It is not the joint distribution of an individual's potential outcomes under both policies. The experiment identifies the two policy-specific distributions separately; it does not reveal how the two unrealized potential outcomes are paired within each individual. Accordingly, within-arm variance and correlation targets should not be interpreted as identifying the variance or full distribution of individual treatment effects. Appendix C defines the distributional statistics used in the learning objective.

A.6 Learned subgroup rules and finite-sample anchors

The preceding results are population identities for fixed subgroup rules. Our implementation learns these rules from data, then freezes them before applying them as routing functions.

To formalize this distinction, let $\mathcal{D}_{\text{con}}$ denote the construction data and write

$$g_t = g_t(\,\cdot\,; \mathcal{D}_{\text{con}}).$$

For an individual independent of the construction data, conditioning on $\mathcal{D}_{\text{con}}$ fixes the rule. The identification argument then applies conditional on those construction data. Although the learned rule is data-dependent, its application to that individual still uses only pretreatment covariates.

This distinction also explains the use of treatment profiles during subgroup construction. Realized intervention histories supply labels for learning a predictor from pretreatment covariates. Once fitted, the predictor assigns both UCG and UTG individuals using the same baseline inputs. Using intervention histories as construction labels is different from selecting a calibration individual according to that individual's realized treatment history. The former produces a pretreatment routing rule; the latter conditions the comparison on post-assignment information.

For an observed sample $\mathcal{S}$, define the number of individuals in each arm of a calibration cell as

$$n_{z,g,t} = \sum_{i \in \mathcal{S}} \mathbf{1}\{Z_i = z, G_{i,t} = g\}.$$

When both counts are positive, the empirical mean-contrast anchor is

$$\widehat{\Delta}_{g,t} = \frac{\sum_{i \in \mathcal{S}} \mathbf{1}\{Z_i = 1, G_{i,t} = g\} V_{i,t}}{n_{1,g,t}} - \frac{\sum_{i \in \mathcal{S}} \mathbf{1}\{Z_i = 0, G_{i,t} = g\} V_{i,t}}{n_{0,g,t}}.$$

This is a finite-sample estimate of the population contrast. Its precision depends on the effective support in both arms and the variability of the quantities being compared. Population comparability does not imply exact covariate balance in every realized sample or an empirical anchor free of sampling error.

Our construction procedure therefore separates causal identification from anchor quality control. Profile purity measures whether a baseline-defined region captures its intended treatment profile. Support, pretreatment balance, and standard-error criteria assess the quality of the observed comparison. These checks determine which cells supply training constraints; they do not establish the causal identity in Proposition 1.

Learning and selection also affect the statistical properties of empirical anchors. A population identity for a fixed rule does not, by itself, guarantee unbiasedness when the same observations are reused to construct or select that rule and estimate its contrast. In particular, selection based on observed outcomes or estimated contrast precision is distinct from the population identification argument. We therefore do not assign a blanket post-selection unbiasedness guarantee to the retained anchors.

Appendix B specifies the construction and reference folds, the frozen routing rules, and the stage-specific acceptance criteria. Appendix C describes how empirical contrasts and distributional targets are estimated and incorporated into training. Together, these procedures use the identified policy contrasts as experimentally grounded supervision while keeping their causal interpretation separate from their finite-sample estimation.

A.7 Identified policy contrasts and replay regularization

Proposition 1 identifies the experimental target $\Delta_{g,t}$; it does not establish equality between that target and the model-side replay contrast in Section 5.4. These are distinct quantities. The target averages trajectories generated by the assigned policies. During training, the model recursively generates states and outcomes while fixing the UTG individual's observed action sequence. During dynamic-policy evaluation, the policy instead selects actions from each generated history.

Fixing observed actions does not generally preserve their dependence on newly generated states. For example, let a post-baseline state satisfy $X_1 \sim \text{Bernoulli}(1/2)$, with policy $A_1 = X_1$ and outcome $Y_2 = A_1 X_1$. The dynamic-policy mean is $1/2$. Replaying an observed A_1 on an independent draw $\widetilde{X}_1$ from the correct state distribution yields $\mathbb{E}[A_1 \widetilde{X}_1] = 1/4$. With a zero-action reference, the same discrepancy holds for the effect. Correct transitions, common initial conditions, and randomized program assignment therefore do not by themselves make the replay contrast equal to the policy contrast.

Accordingly, UHC uses experimental anchors as surrogate regularization targets alongside factual learning, free-rollout supervision, and within-arm distributional calibration. The identification result establishes the causal meaning of the targets, not consistency of the replay-based learning objective for dynamic-policy effects. We assess the resulting dynamic predictions directly: Appendix H.1 reports their subgroup-average errors against simulator references on an independent test cohort. This comparison evaluates integrated UHC; it neither isolates the contrast loss nor establishes replay–policy equivalence.

B Subgroup Construction and Anchor Selection

Universal Holdout Calibration accepts fixed subgroup rules defined by pretreatment information. Our implementation constructs these rules in two stages: it first identifies baseline regions associated with distinct intervention profiles, then subdivides the retained regions using initial states. The first stage separates expected intervention patterns; the second distinguishes starting conditions within those regions. Both stages produce routing functions that are applied identically to UCG and UTG.

This appendix describes the complete construction procedure, including the different acceptance criteria applied before and after initial-state refinement. As established in Appendix A, these criteria assess the usefulness and finite-sample quality of calibration cells. They are separate from the randomized-assignment argument that identifies their population policy contrasts.

B.1 Data partitioning and construction roles

We partition the source cohort of 90,000 individuals into three disjoint folds, denoted P, A, and M. Splitting is stratified by randomized arm, yielding 30,000 individuals per fold: 15,000 UCG and 15,000 UTG individuals. The reported experiments use construction seed 2032 and role rotation 0; the corresponding role-splitting seed is $2032 + 31{,}337 = 33{,}369$.

Fold P supplies the data used to learn subgroup routing rules. Its UTG trajectories are used to construct treatment-profile labels and fit their pretreatment predictors. Both arms of Fold P are used for the subsequent initial-state refinement. Fold A supplies the reference observations used to assess candidate cells and construct within-arm distributional targets. Fold M denotes the remaining third of the source cohort.

These roles organize subgroup construction, not the final model-training population. The trajectory models are trained using all 90,000 source individuals, including those in P and A. In particular, model training is not restricted to Fold M. Validation and test cohorts do not participate in subgroup construction.

B.2 Constructing time-specific treatment profiles

For each target time $t \in \{1, \ldots, T\}$, we construct treatment-profile features from the observed UTG dose prefix preceding that time,

$$A_{i,0:t-1}.$$

In CVSim, this prefix contains fluid and vasopressor doses. Its features summarize intervention family, treatment frequency, episode counts, timing, cumulative and recent doses, dose variation, and transitions between treatment states. The profile at time t uses no interventions after $t-1$. Outcomes and physiological states are not inputs to this profile-construction step.

Profiles are constructed separately at each time using the 15,000 UTG individuals in Fold P. Each feature is centered by its median and scaled by its interquartile range. A degenerate interquartile range is replaced by the feature's standard deviation, with a unit scale used if both are degenerate. Standardized values are clipped to $[-8, 8]$, and effectively constant columns are removed before clustering.

We use MiniBatch K-means to identify profile classes. The effective maximum number of profile clusters is four. Candidate solutions are examined in descending order, $K = 4, 3, 2$, and the largest admissible solution is retained. Every cluster must contain at least

$$\max(100,\ \lceil 0.02\, N_{P,\mathrm{UTG}} \rceil) = 300$$

individuals. Cluster centers must also satisfy a minimum separation requirement. For centers c_k, c_ℓ in the d_{active}-dimensional standardized feature space, the criterion is

$$\min_{k \neq \ell} \frac{\|c_k - c_\ell\|_2}{\sqrt{d_{\text{active}}}} \geq 0.2.$$

Clustering uses k-means++ initialization, five initializations, at most 200 iterations, and mini-batches of at most 2,048 individuals. Cluster identifiers are ordered deterministically, and membership is recomputed from the stored centers. Support requirements are checked again after this reassignment. If no multi-cluster solution satisfies the requirements, the time point receives an unsplit population label; this label is not used as a proper calibration subgroup.

The resulting labels describe realized treatment profiles among Fold-P UTG individuals. They are training targets for the pretreatment filter described next, not the final rule used to assign UCG and UTG individuals to calibration groups.

B.3 Learning a common pretreatment filter

At each time, we train a Random Forest classifier to predict the treatment-profile label from pretreatment covariates. The classifier is fitted on Fold-P UTG individuals. Its inputs are the 18 observed baseline measurements: the 17 initial-state variables and the initial MAP outcome. Hidden simulator parameters and post-intervention measurements are excluded.

Baseline features are standardized using means and standard deviations estimated from Fold-P UTG. Each forest contains 96 trees, with maximum depth 10 and a minimum of 25 samples per leaf. Training uses bootstrap sampling, out-of-bag predictions, and class-balanced subsample weights.

We use out-of-bag predictions to determine a separate confidence threshold for each profile class at each time. Let $\widehat{C}_{i,t}^{\mathrm{OOB}}$ denote the predicted class and $q_{i,t}^{\mathrm{OOB}}$ its predicted probability. For class c and candidate threshold q, define

$$\mathcal{S}_{c,t}(q) = \left\{ i : \widehat{C}_{i,t}^{\mathrm{OOB}} = c,\ q_{i,t}^{\mathrm{OOB}} \geq q \right\}.$$

Candidate thresholds are examined in ascending order. We select the smallest threshold for which the accepted set contains at least 50 individuals and its empirical precision is at least 0.6:

$$|\mathcal{S}_{c,t}(q)| \geq 50, \qquad \frac{\sum_{i \in \mathcal{S}_{c,t}(q)} \mathbf{1}\{C_{i,t} = c\}}{|\mathcal{S}_{c,t}(q)|} \geq 0.6.$$

Here, $C_{i,t}$ is the treatment-profile label obtained from the observed dose prefix. The value 0.6 is therefore a target precision among accepted predictions, not a common probability cutoff imposed on every class.

After fitting, the classifier assigns an individual to its highest-probability class only when that class has a valid threshold and the prediction meets it. Otherwise, the individual is assigned to an unknown region. The fitted feature transformation, forest, and thresholds are frozen and applied identically to both randomized arms.

Consequently, UCG individuals receive profile-based assignments without supplying an intervention history from the treatment program. Their membership is determined entirely by baseline information. Treatment histories are used to learn what the baseline filter predicts; they are not required when applying that filter.

B.4 Selecting parent calibration cells

We apply the frozen filter to Fold A and evaluate each candidate parent subgroup–time cell. Membership is determined by the pretreatment filter. For UTG individuals, we additionally obtain their realized profile labels by applying the frozen profile partition to their observed dose prefixes. These labels are used to evaluate profile purity.

For parent class c at time t, purity is the fraction of UTG individuals routed to that class whose realized profile label agrees:

$$\mathrm{Purity}_{c,t} = \frac{\sum_{i \in A} \mathbf{1}\{Z_i = 1, R_t(B_i) = c\} \mathbf{1}\{C_{i,t} = c\}}{\sum_{i \in A} \mathbf{1}\{Z_i = 1, R_t(B_i) = c\}},$$

where R_t denotes the frozen parent-routing function. This Fold-A purity assessment is distinct from the out-of-bag precision criterion used to choose classifier thresholds.

A parent cell must satisfy the following requirements:

- At least 100 UCG and 100 UTG individuals.
- Profile purity of at least 0.6.
- Maximum absolute standardized mean difference across the 18 baseline features of at most 0.1.

- Estimated standard error of the observed outcome contrast of at most 5 mmHg.

For a baseline feature b_j, let $\bar{b}_{z,j}$, $s^2_{z,j}$, and n_z denote its arm-specific mean, sample variance, and sample size within the cell. We calculate the pooled variance as

$$s^2_{p,j} = \frac{(n_1 - 1)s^2_{1,j} + (n_0 - 1)s^2_{0,j}}{n_1 + n_0 - 2},$$

and define the balance statistic as

$$\text{SMD}_{\max} = \max_j \frac{|\bar{b}_{1,j} - \bar{b}_{0,j}|}{s_{p,j}}.$$

For a pooled standard deviation at or below 10^{-12}, the component is assigned zero if its absolute mean difference is also at or below 10^{-12}, and infinity otherwise.

The estimated standard error of the outcome contrast is

$$\widehat{\text{SE}}_{c,t} = \sqrt{\frac{s^2_{Y,1}}{n_1} + \frac{s^2_{Y,0}}{n_0}},$$

using sample variances of the observed Y_t values in the two arms. This calculation uses the original outcome scale, so the threshold of 5 corresponds to 5 mmHg for MAP.

We further restrict calibration to proper, informative profile cells. A proper cell contains a strict subset of the Fold-A UTG population; unknown regions and population-fallback cells are excluded. An informative cell contains at least one UTG individual with at least one non-no-op intervention in the prefix preceding t.

The informative-prefix requirement does not require every UTG individual in the cell to receive treatment. Untreated UTG individuals remain members of their baseline-defined subgroup. Likewise, rejecting a cell removes a calibration constraint, not its individuals from ordinary trajectory training.

B.5 Refinement by initial state

For each retained parent cell, we subdivide its baseline region using the initial state X_0. This second stage separates starting conditions within a region already associated with a common treatment profile.

The refinement is fitted on all Fold-P individuals routed to the parent cell, pooling UCG and UTG. Its inputs are the 17 initial-state variables. Initial MAP, arm labels, outcomes, actions, and doses are not inputs to the refinement algorithm. Thus, the use of initial MAP in the parent classifier is distinct from the state-only input used for this second stage.

Within each parent–time region, we center initial-state features by their medians and scale them by their interquartile ranges. Degenerate scales use the same standard-deviation and unit-scale fallbacks described above. Values are clipped to $[-8, 8]$. MiniBatch K-means then partitions the standardized initial states into exactly two child groups. We use k-means++ initialization, five initializations, at most 200 iterations, mini-batches of at most 2,048 individuals, and no cluster reassignment.

The fitted transformations and centers are frozen. For an individual routed to an accepted parent c, child membership is determined by the nearest stored center:

$$k_{c,t}(X_{i,0}) = \arg\min_{k \in \{1,2\}} \left\| \widetilde{X}^{(c,t)}_{i,0} - m_{c,t,k} \right\|_2^2,$$

where $\widetilde{X}^{(c,t)}_{i,0}$ is the parent-specific transformed initial state and $m_{c,t,k}$ is the corresponding center. The final subgroup is identified by the parent–child pair

$$G_{i,t} = \left(R_t(B_i),\, k_{R_t(B_i),t}(X_{i,0})\right).$$

Within each retained parent, the two children are disjoint and exhaust the parent's membership. The same routing operation is used for both arms.

B.6 Validating refined cells

After fitting the refinement on Fold P, we apply it to Fold A and evaluate every child cell. The required conditions are:

- At least 100 individuals in each arm.
- Maximum absolute SMD across the 17 initial-state variables of at most 0.1.
- Estimated standard error of the observed outcome contrast of at most 5 mmHg.
- At least one non-no-op intervention in the observed prefix of at least one UTG member.

The balance and standard-error calculations follow the definitions above. The difference is the feature set used for balance: parent cells are checked on all 18 baseline measurements, whereas child cells are checked on the 17 initial-state variables.

Child profile purity is recorded as a diagnostic but is not an acceptance criterion. The parent filter has already established the connection to the intended treatment profile. The purpose of refinement is to distinguish initial states within that region, and the implementation does not require every resulting child to satisfy the parent-level purity threshold.

The construction enforces the configured two-child partition as a whole. If any child fails a required support, balance, precision, or informative-prefix condition, construction terminates with an error. The procedure does not silently discard that child, retain only its sibling, or reduce the number of clusters. This preserves the specified nested partition rather than changing its structure through selective child removal.

B.7 Final calibration cells and training interface

After construction, the parent classifiers, confidence thresholds, initial-state transformations, child centers, and accepted cell identifiers are frozen. They are applied to the full source cohort to determine calibration membership. Because the construction is time-specific, an individual can belong to different subgroups at different prediction times. Every assignment nevertheless remains a function of pretreatment information.

For the reported experiment, the procedure retains 13 parent subgroup–time cells and produces 26 child subgroup–time cells. These occur at target times

$$t \in \{1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 14, 15, 16\}.$$

Each retained parent contributes two children. Separating UCG and UTG within the 26 cells gives 52 subgroup–time–arm combinations for distributional calibration. Each combination supplies multiple distributional statistics; 52 is not the number of individual scalar loss targets.

In Fold A, the retained child cells contain between 1,170 and 5,443 UTG individuals and between 1,102 and 5,467 UCG individuals. Their outcome-contrast standard errors range from 0.1702 to 1.0932 mmHg, and their maximum absolute initial-state SMDs range from 0.0296 to 0.0707. Diagnostic child profile purities range from 0.4710 to 0.9583, reflecting their diagnostic-only role after refinement.

These support counts and precision summaries describe the reference-fold cells. They are distinct from the effective arm counts and estimation noise within a training mini-batch. The frozen cells define where calibration is applied; Appendix C specifies how they are used to estimate mini-batch contrasts, construct within-arm distributional targets, and form the integrated learning objective.

C Calibration Objectives and Mini-batch Estimation

This appendix specifies the GT calibration objective, including target construction, normalization, sampling, and optimization. Backbone-specific adaptations and intermediate latent supervision are described in Appendix D.

C.1 Factual and free-rollout prediction

The state contains 16 continuous variables and binary PE; the outcome is MAP. Continuous variables use a common normalization fitted on the source-training cohort.

Let e_X denote the sum of squared errors over the 16 continuous states, e_Y the squared outcome error, and ℓ_{PE} the binary-state cross-entropy. Factual prediction uses observed histories and minimizes

$$\mathcal{L}_{\text{factual}} = \left\langle \ell_{\text{PE}} + \frac{e_X + e_Y}{17} \right\rangle,$$

where brackets denote averaging over valid patient–time observations.

Free rollouts start from observed initial conditions and recursively feed back predicted categorical probabilities and continuous conditional means, without residual sampling. UCG trajectories follow no-op actions; UTG trajectories replay their observed intervention sequences. Each arm is normalized separately:

$$\mathcal{L}^{z}_{\text{free}} = \left\langle \frac{e_X + \ell_{\text{PE}}}{17} + e_Y \right\rangle_{Z=z}, \qquad z \in \{0, 1\}.$$

Here, errors are computed from free-rollout predictions. Unlike the factual objective, the free-rollout objective gives the outcome its own full-weight term.

Training rollouts replay fixed observed actions. They do not invoke the dynamic policy to choose new actions from predicted histories; that operation belongs to counterfactual evaluation.

C.2 Mini-batch mean-contrast calibration

For each accepted subgroup–time cell represented in both arms of a training batch $\mathcal{B}$, the observed anchor and model-side contrast are

$$\widehat{\Delta}^{\mathcal{B}}_{g,t} = \overline{V}^{\mathcal{B}}_{1,g,t} - \overline{V}^{\mathcal{B}}_{0,g,t}, \qquad \widehat{d}^{\mathcal{B}}_{\theta,g,t} = \frac{1}{n^{\mathcal{B}}_{1,g,t}} \sum_{i \in \mathcal{I}^{\mathcal{B}}_{1,g,t}} \left(\widetilde{V}^{\text{rep}}_{i,t} - \widetilde{V}^{0}_{i,t} \right).$$

The two predicted trajectories start from the same UTG individual's initial conditions. The teacher is an observed UTG–UCG difference, not a difference involving UCG model predictions. Observed targets are detached; both predicted branches retain gradients.

The contrast loss is

$$\mathcal{L}_{\text{contrast}} = \sum_{(g,t) \in \mathcal{C}_{\mathcal{B}}} w_{g,t} \left[\left(\widehat{d}^{Y}_{\theta,g,t} - \widehat{\Delta}^{Y}_{g,t} \right)^2 + \frac{1}{17} \left\| \widehat{d}^{X}_{\theta,g,t} - \widehat{\Delta}^{X}_{g,t} \right\|_2^2 \right].$$

The state vector includes PE, represented by its observed indicator and predicted probability. Continuous components use the shared training normalization.

A cell is eligible when the batch contains at least one valid individual from each arm. Missing-arm cells are skipped. With K_t denoting the number of frozen cells at time t, weights are proportional to $1/K_t$ and renormalized over eligible cells. The construction-time support requirement of 100 individuals per arm does not apply within each training batch.

C.3 Fixed distributional targets and losses

For each subgroup–time–arm combination, we compute fixed targets from Fold-A observations of the 16 normalized continuous states and normalized outcome. Binary PE is excluded. The statistics are:

- **Variance:** empirical variance with divisor N.
- **Quantiles:** linearly interpolated empirical quantiles at probabilities $0.1, 0.5, 0.9$.
- **Tails:** lower and upper partial moments and smooth exceedance probabilities relative to fixed observed 5th and 95th percentiles.
- **Correlation:** contemporaneous X–X and X–Y correlations.

For an observed lower threshold l, upper threshold u, and variable W, the four tail statistics are

$$\mathbb{E}[(l-W)_+], \quad \mathbb{E}[(W-u)_+], \quad \mathbb{E}\left[\sigma\left(\frac{l-W}{h}\right)\right], \quad \mathbb{E}\left[\sigma\left(\frac{W-u}{h}\right)\right],$$

with $h = 0.1 \max(\mathrm{SD}_{\mathrm{obs}}, 0.25)$. Predicted statistics use the same fixed thresholds and bandwidths.

All distributional comparisons use Smooth-L1 loss with transition parameter 1. Variances are compared after transformation by $\log(v + 0.05)$. Quantiles and tail partial moments are divided by $\max(\mathrm{SD}_{\mathrm{obs}}, 0.25)$; smooth tail probabilities require no scale adjustment. Partial-moment and probability losses are added.

For variance, quantile, and tail terms, aggregation sums the mean outcome error and the mean continuous-state error. Correlation loss averages over valid upper-triangular variable pairs. Both variables must have observed variance above 10^{-6}; correlation denominators use the same variance floor. Covariances use divisor N. No separate covariance or cross-time correlation loss is included.

The combined distributional loss is

$$\mathcal{L}_{\mathrm{distribution}} = \frac{1}{4}\left(\mathcal{L}_{\mathrm{variance}} + \mathcal{L}_{\mathrm{quantile}} + \mathcal{L}_{\mathrm{tail}} + \mathcal{L}_{\mathrm{correlation}}\right).$$

C.4 Support, sampling, and stochastic prediction

Fixed distributional targets require at least 128 distinct observed patients per cell–arm; tail targets require 256. Their respective weights are

$$\omega_{z,g,t} = \frac{n_{z,g,t}}{n_{z,g,t} + 128} \frac{1}{K_t}, \qquad \omega^{\mathrm{tail}}_{z,g,t} = \frac{n_{z,g,t}}{n_{z,g,t} + 256} \frac{1}{K_t}.$$

Each loss component is normalized by the sum of its contributing weights. Predicted statistics require at least 16 distinct patients per arm.

At each optimizer step, we additionally sample one cell supported in both arms, with probability proportional to the sum of its arm weights. From that Fold-A cell, we uniformly sample 64 distinct individuals per arm without replacement and generate two stochastic replay trajectories per individual. Only the sampled cell is scored.

Statistics are computed by pooling patients and particles within each arm, not by first averaging particles within each patient. Thus, the predicted statistics retain both between-patient variation and stochastic trajectory variation. Particles do not increase observed support counts.

Stochastic replay uses hard categorical draws and continuous conditional means plus marginally sampled empirical residuals. Categorical draws and residual samples are detached; gradients propagate through continuous predictions and their recurrent histories. The residual bank is estimated from 512 fixed Fold-A individuals, initialized before training, and refreshed after each epoch. This updates the prediction-sampling mechanism, not the fixed observed distributional targets.

C.5 Joint objective, updates, and anchor precision

For epoch e, define $r_k(e) = \min(1, e/k)$. The common calibration objective is

$$\boxed{\begin{aligned}\mathcal{L}_{\text{UHC}}(e) = \mathcal{L}_{\text{factual}} + 0.1\, r_5(e)\left(\mathcal{L}^0_{\text{free}} + \mathcal{L}^1_{\text{free}} + \mathcal{L}_{\text{contrast}}\right) \\ + 0.1\, r_3(e)\mathcal{L}_{\text{distribution}}.\end{aligned}}$$

Factual, free-rollout, and contrast losses use batches of 512 individuals from the full source cohort. Distributional calibration uses the additional Fold-A sample described above. We accumulate gradients from these objectives, clip the total gradient norm to 1, and perform one optimizer update. Factual prediction uses training-mode dropout; calibration rollouts disable dropout while retaining gradient computation. Staged latent supervision adds the intermediate objectives specified in Appendix D.

The batch size of 512 is not a per-cell support count. Mean-contrast anchor precision depends on the actual arm counts and outcome variability within each batch cell. These fluctuations are distinct from both the reference-fold standard errors reported in Appendix B and the Monte Carlo variability of predicted distributional statistics. Reference-fold precision therefore does not directly quantify mini-batch anchor noise.

D Model Architectures and Task Adaptations

G-Transformer and comparison variants. GT uses separate categorical and continuous encoders, each with three layers, hidden width 128, eight attention heads, feedforward width 512, and dropout 0.1. Each transition is predicted from its available history prefix, preserving the backbone's categorical–continuous conditioning structure. Pooled GT, Anchor GT, and Calibrated GT share this architecture. Loop GT repeats each block four times, sharing weights across repetitions of that block while retaining distinct weights between blocks. Latent CoT instead performs three passes with two latent-feedback updates.

Latent-feedback recurrence in GT. Each encoder shares its own parameters across passes. Let $b \in \{\text{cat}, \text{cont}\}$ index the encoder and let $E^{(0)}_{b,t}$ denote its original embedded input for the transition at time t. Suppressing the individual index, the representation used for prediction after pass $r \in \{0, 1, 2\}$ is

$$h^{(r)}_{b,t} = \left[F_{\theta_b}\left(E^{(r)}_{b,t}\right)\right]_{q_t},$$

where q_t is the original history position used for next-step prediction.

For $r = 0, 1$, this representation is converted into a feedback token:

$$z^{(r)}_{b,t} = \phi\left(h^{(r)}_{b,t} + e_r\right),$$

and the next pass receives

$$E^{(r+1)}_{b,t} = \left[E^{(0)}_{b,t}; z^{(0)}_{b,t}; \ldots; z^{(r)}_{b,t}\right].$$

Here, ϕ is a learned feedback MLP shared across both encoders and recurrences, e_r is a learned pass embedding, and concatenation is along the token dimension. Each pass therefore accesses the original input together with all feedback tokens produced so far. The prediction readout remains at the original query position. Feedback contains latent representations, not decoded state or outcome predictions.

Staged latent supervision. The staged variant retains the ordinary Latent CoT architecture and adds state supervision after pass 0 and outcome supervision after pass 1, with coefficients 0.1 and 0.2. Let $\mathcal{P}$ denote the set of valid observed history–action prefixes. For each $(i, t) \in \mathcal{P}$, the targets are the next state $X_{i,t+1}$ and outcome $Y_{i,t+1}$. The early state objective is

$$\mathcal{L}^{(0)}_{X,\text{factual}} = \frac{1}{|\mathcal{P}|} \sum_{(i,t)\in\mathcal{P}} \ell_X\left(\widehat{X}^{(0)}_{i,t+1}, X_{i,t+1}\right).$$

The loss ℓ_X combines squared prediction errors for continuous state variables with cross-entropy for categorical states. The intermediate outcome objective is

$$\mathcal{L}^{(1)}_{Y,\text{factual}} = \frac{1}{|\mathcal{P}|} \sum_{(i,t)\in\mathcal{P}} \ell_Y \left(\widehat{Y}^{(1)}_{i,t+1}, Y_{i,t+1} \right),$$

where ℓ_Y is the normalized squared-error component of the factual outcome objective.

Both losses preserve the categorical–continuous weighting and dimensional normalization of the backbone's factual objective, as defined in Appendix C. Existing prediction heads are reused to read out intermediate representations; no additional stage-specific prediction heads are introduced. The auxiliary losses are evaluated only on observed prefixes, not on separate intermediate counterfactual rollouts. Their supervision comes entirely from observed states and outcomes, without additional individual counterfactual labels.

Pass 2 supplies the final joint predictions used by the complete UHC objective. Effects are computed from differences between final-pass rollout outcomes, not from a separate effect head. Calibrated GT and Loop GT each contain 1,198,739 parameters; unstaged and staged Latent CoT each contain 1,215,827. The latter pair also shares inference computation, isolating the addition of intermediate supervision.

CT and Mamba-CDSP adaptations. Both backbones receive a next-state prediction head alongside outcome prediction and represent interventions using three classes—no-op, fluid, and vasopressor—and two continuous doses. Predicted states and outcomes are returned to the history for recursive dynamic-policy evaluation. Pooled training includes UCG trajectories as factual no-op observations; randomized-arm labels are not model inputs. CT retains its three-stream attention and treatment-balancing mechanism, using hidden width 48, one layer, two heads, and representation/head widths of 24. Our paper-based Mamba-CDSP reimplementation retains selective state-space modeling and CDSP regularization, using model width 32, SSM state dimension 16, and one block. The proposed CT extension adds the same UHC configuration and three-pass staged supervision, with auxiliary coefficients 0.1 and 0.2 applied to CT's native state and outcome losses. Its feedback MLP is shared across all three streams and both recurrences. Unlike GT, CT's causal attention requires readout from the last original history position on pass 0 and the newest feedback token thereafter. The current intervention conditions the prediction heads but does not enter latent feedback.

E Benchmark Construction and Data Quality Control

Simulation and interventions. We use CVSim (Heldt et al., 2010) to generate longitudinal trajectories with 17 observed state variables—16 continuous variables and binary pulmonary embolism status—and mean arterial pressure (MAP) as the outcome. Following a 30-second warm-up, observations are recorded at one-second intervals over 16 prediction steps. Interventions comprise no-op, fluid administration, and vasopressor administration, with continuous doses for the two active treatments. UTG patients follow a dynamic treatment policy, whereas UCG patients receive no intervention from the target program. Patient-specific simulator parameters and latent treatment-response factors are excluded from model inputs.

Benchmark configuration. Relative to the default implementation, we increase the treatment-effect scale from 1.0 to 1.25 and the policy's MAP target from 70 to 90 mmHg. We additionally require baseline MAP $\leq$ 85 mmHg and central venous pressure $\leq$ 18 mmHg. Under this configuration, 81.9% of source-cohort UTG patients receive at least one intervention during follow-up. This is patient-level treatment coverage, not the proportion of treated time steps. These settings produce substantial policy–no-op contrasts, avoiding a benchmark dominated by near-zero effects.

We introduce unobserved response heterogeneity through patient-specific, mean-one lognormal multipliers:

$$\log R \sim \mathcal{N}(-\sigma^2/2, \sigma^2).$$

The log-scale standard deviations are 0.25 for overall treatment response and 0.35 each for fluid-specific and vasopressor-specific responses, compared with zero in the default configuration. These multipliers remain fixed within each patient. We set the dose-noise scale to 0.1, the probability of refreshing visit-level policy randomness to 0.05, and the disease-event probability to 0.0015. Otherwise, policy decisions reuse patient-specific latent propensities.

Cohort construction and quality control. Baseline screening excludes candidates exceeding the candidate-pool 99th percentile in any recorded baseline variable, together with candidates failing the baseline eligibility criteria. We then simulate both policy and no-op paths and require finite measurements, MAP ≤ 250 mmHg, and heart rate ≤ 250 beats per minute at every recorded time under both paths. Randomized arm assignment occurs after screening. The source cohort contains 90,000 patients, equally divided between UCG and UTG; model training receives only each patient's assigned-arm trajectory. We use 1,000 validation patients and an independent test cohort initially containing 6,000 patients. Source and validation generation use seed 2032; independent test generation uses seed 2036.

After constructing the test reference trajectories described in Appendix G, an additional, post-hoc quality-control screen excludes three patients whose saved Monte Carlo mean MAP exceeds 250 mmHg under either policy. This screen depends on reference trajectories, not model predictions or prediction errors. All models are evaluated on the same remaining 5,997 patients.

F Training Conditions

All models use Adam with zero weight decay, gradient-norm clipping at 1.0, and an 80-epoch limit. The main comparisons use training seed 2032; Appendix I additionally evaluates Calibrated GT and its unstaged and staged Latent CoT extensions with training seeds 2031 and 2033. Continuous inputs and targets use normalization fitted on the source-training cohort.

GT variants use batch size 512 and an initial learning rate of 10^{-4}, with cosine-annealing warm restarts every 10 epochs and a minimum learning rate of 10^{-5}. Pooled CT and Calibrated CT with Staged Latent CoT both use batch size 512 and fixed learning rates of 10^{-3} for the prediction model and treatment classifier, with alternating updates. Mamba-CDSP uses batch size 128, a fixed learning rate of 10^{-2}, and an exponential moving average of model parameters with decay 0.99. Architecture settings and auxiliary-loss coefficients are specified in Appendix D; calibration-loss weights and warm-up schedules are specified in Appendix C.

Checkpoints are selected using factual prediction loss on the observed validation cohort. GT and CT use final-output factual validation prediction loss, and the two CT variants share their optimizer hyperparameters and checkpoint-selection criterion. The common epoch limit does not equate computation across backbones or recurrent variants.

G Counterfactual Evaluation

Reference trajectories and model rollouts. For each test patient, we construct policy and no-op reference trajectories by averaging eight simulator realizations. Patient-specific latent physiology and response factors remain fixed, while future stochastic events are averaged. The reference effect at each time is the difference between the two reference mean outcomes. These counterfactual references are used only for evaluation.

Both model rollouts start from the same observed initial state and outcome. The dynamic policy selects interventions from the model-generated history; the reference branch fixes intervention type and dose to zero. Neither branch receives future observed states, outcomes, or realized intervention sequences. Effect predictions are the difference between the predicted policy and no-op outcomes.

Model evaluation reuses the eight saved policy-randomness realizations. For GT, we generate 100 stochastic trajectories per patient and branch, distributed across these eight realizations. Predictions are averaged within each realization and then equally across realizations. CT and Mamba-CDSP recursively use deterministic transition predictions, generating eight policy trajectories and one deterministic no-op trajectory per patient. The eight policy trajectories are averaged equally. Sharing policy randomness does not impose identical intervention sequences: each model's generated history determines its subsequent interventions.

Prediction metrics. Let $\widetilde{\mu}^{\rho}_{i,t}$ denote the Monte Carlo reference expected outcome for patient i under policy ρ, and let

$$\widetilde{\tau}_{i,t} = \widetilde{\mu}^{\pi}_{i,t} - \widetilde{\mu}^{\pi_{\text{ref}}}_{i,t}.$$

For n test patients and prediction horizon T, the primary metrics are

$$\text{Effect RMSE} = \left[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\widehat{\tau}_{i,t} - \widetilde{\tau}_{i,t}\right)^2\right]^{1/2},$$

and

$$\text{Policy Y RMSE} = \left[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\widehat{\mu}^{\pi}_{i,t} - \widetilde{\mu}^{\pi}_{i,t}\right)^2\right]^{1/2}.$$

The first measures the accuracy of individual-specific expected policy effects throughout the trajectory; the second measures expected outcome prediction under the dynamic policy. Terminal Effect RMSE restricts the effect comparison to $T = 16$; No-op Y RMSE applies the trajectory-level outcome metric to the reference policy. Terminal Pearson correlation measures the linear association between predicted and reference terminal effects.

One-step state prediction is evaluated on observed histories using the 16 continuous state variables, excluding binary PE and the MAP outcome:

$$\text{X NRMSE} = \sqrt{\frac{1}{16nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{16}\left(\frac{\widehat{X}^{\text{1step}}_{i,t,j} - X_{i,t,j}}{s_j}\right)^2},$$

where s_j is the source-training standard deviation of variable j. This metric evaluates one-step factual prediction, not counterfactual rollout accuracy.

Effect-ranking metrics. Patients are ranked by predicted terminal effect in descending order. Writing $\widetilde{\tau}_{(j),T}$ for the reference effect of the patient at rank j, the cumulative gain curve is

$$G(k/n) = \frac{1}{n}\sum_{j=1}^{k}\widetilde{\tau}_{(j),T}, \qquad G(0) = 0.$$

Exact prediction ties are averaged. Qini is the trapezoidal area under this curve minus the random-ranking area:

$$\text{Qini} = \text{AUC}(G) - \frac{\bar{\bar{\tau}}_T}{2}.$$

Normalized AUUC is the ratio of this baseline-corrected area to the corresponding oracle value, obtained by ranking patients by their reference effects:

$$\text{Normalized AUUC} = \frac{\text{Qini}}{\text{Qini}_{\text{oracle}}}.$$

Paired patient-bootstrap evaluation. We quantify uncertainty using 100,000 paired patient-bootstrap resamples. Each sampled patient contributes the complete trajectory, and the same resample is used for all models. Model-specific intervals are pointwise 95% confidence intervals. Comparisons between the proposed GT model and the six alternatives use Bonferroni correction separately within each primary metric; Table H.4 reports these adjusted intervals. The CT transfer comparison uses the same patient-bootstrap procedure and is reported separately. The supervision-schedule ablation forms a separate family of 36 comparisons: two alternative schedules across all 18 evaluated metrics (Appendix H.3). These intervals quantify patient-sampling uncertainty conditional on the fitted checkpoints and saved Monte Carlo estimates; they do not incorporate training-seed variation or additional Monte Carlo estimation uncertainty.

H Supplementary Results

Tables H.1 and H.2 report supplementary metrics; Tables H.3 and H.4 report paired differences; and Tables H.5 and H.6 report terminal Effect RMSE. The seven-model main comparison and separate CT transfer comparison remain distinct. Model-specific intervals are pointwise 95% paired patient-bootstrap confidence intervals. In model-metric tables, bold denotes the best point estimate within each comparison.

Direct comparisons. Table H.3 complements the Stage-versus-baseline comparisons in Table H.4 by isolating calibration, ordinary looping, and unstaged latent computation within GT. The CT comparison evaluates the combined calibration and staged-supervision method.

Table H.1: Supplementary metrics for the seven-model comparison. Pearson correlation, Qini, and normalized AUUC evaluate terminal effects. No-op Y RMSE evaluates the outcome trajectory; X NRMSE evaluates one-step state prediction.

Model	Pearson ↑	Qini ↑	Normalized AUUC ↑	No-op Y RMSE ↓	X NRMSE ↓
Pooled GT	0.6970 [0.6818, 0.7117]	3.1592 [3.0592, 3.2596]	0.7393 [0.7260, 0.7522]	6.6894 [6.5593, 6.8345]	0.182493 [0.175323, 0.190498]
Calibrated GT	0.8438 [0.8341, 0.8531]	3.7209 [3.6241, 3.8185]	0.8707 [0.8629, 0.8782]	1.4926 [1.2894, 1.8025]	0.132227 [0.128851, 0.136265]
Calibrated GT + Loop	0.8406 [0.8307, 0.8502]	3.7203 [3.6240, 3.8172]	0.8706 [0.8629, 0.8779]	**1.4177** [1.2055, 1.7393]	0.121407 [0.117531, 0.126027]
Calibrated GT + Latent CoT	0.8404 [0.8305, 0.8500]	3.7179 [3.6214, 3.8149]	0.8700 [0.8621, 0.8775]	1.4420 [1.2348, 1.7579]	0.123874 [0.120124, 0.128367]
Calibrated GT + Staged Latent CoT	**0.8455** [0.8358, 0.8550]	**3.7341** [3.6376, 3.8313]	**0.8738** [0.8660, 0.8812]	1.4458 [1.2380, 1.7612]	0.124130 [0.120512, 0.128456]
Pooled CT	0.8023 [0.7912, 0.8133]	3.5696 [3.4738, 3.6662]	0.8353 [0.8262, 0.8442]	3.3137 [3.1934, 3.4966]	0.123073 [0.121032, 0.125208]
Pooled Mamba-CDSP	0.7659 [0.7529, 0.7787]	3.4136 [3.3139, 3.5134]	0.7988 [0.7869, 0.8104]	3.2050 [3.0547, 3.3890]	**0.120443** [0.118414, 0.122570]

Table H.2: Supplementary metrics for transfer to CT.

Model	Pearson ↑	Qini ↑	Normalized AUUC ↑	No-op Y RMSE ↓	X NRMSE ↓
Pooled CT	0.8023 [0.7912, 0.8133]	3.5696 [3.4738, 3.6662]	0.8353 [0.8262, 0.8442]	3.3137 [3.1934, 3.4966]	**0.123073** [0.121032, 0.125208]
Calibrated CT + Staged Latent CoT	**0.8478** [0.8378, 0.8576]	**3.7396** [3.6428, 3.8367]	**0.8751** [0.8674, 0.8826]	**1.4068** [1.1970, 1.7225]	0.144174 [0.141292, 0.147424]

Table H.3: Differences in primary metrics. Each difference is the first model minus the second; negative values favor the first model. Brackets contain pointwise paired 95% confidence intervals. The multiplicity-adjusted Stage GT comparisons are reported separately in Table H.4.

Comparison	Effect RMSE difference	Policy Y RMSE difference
Calibrated GT − Pooled GT	−4.3839 [−4.6410, −4.1281]	−2.7692 [−2.9654, −2.5736]
Calibrated GT + Loop − Calibrated GT	−0.0376 [−0.1101, 0.0367]	0.0144 [−0.0563, 0.0869]
Calibrated GT + Latent CoT − Calibrated GT	0.0423 [−0.0324, 0.1212]	0.0630 [−0.0127, 0.1424]
Calibrated CT + Staged Latent CoT − Pooled CT	−1.4712 [−1.6784, −1.2642]	−2.1842 [−2.3966, −1.9713]

Table H.4: Paired differences in the two primary metrics. Differences are computed as Calibrated GT + Staged Latent CoT minus the comparison model; negative values favor the proposed model. Intervals are Bonferroni-adjusted to provide 95% familywise coverage over the six comparisons separately for each metric.

Comparison model	**Effect RMSE difference [adjusted CI]**	**Policy Y RMSE difference [adjusted CI]**
Pooled GT	−4.6468 [−5.0159, −4.2782]	−2.9642 [−3.2447, −2.6853]
Calibrated GT	−0.2629 [−0.3735, −0.1469]	−0.1950 [−0.3039, −0.0820]
Calibrated GT + Loop	−0.2253 [−0.3229, −0.1294]	−0.2095 [−0.3052, −0.1152]
Calibrated GT + Latent CoT	−0.3052 [−0.3923, −0.2212]	−0.2580 [−0.3418, −0.1768]
Pooled CT	−1.3432 [−1.6364, −1.0485]	−2.0697 [−2.3723, −1.7610]
Pooled Mamba-CDSP	−1.0435 [−1.2890, −0.7995]	−1.0742 [−1.3276, −0.8233]

Table H.5: Terminal effect prediction for the seven-model main comparison. Entries retain the point estimates and pointwise 95% paired patient-bootstrap confidence intervals from the main comparison.

Model	**Effect RMSE,** $t = 16 \downarrow$
Pooled GT	15.9973 [15.6402, 16.3610]
Calibrated GT	8.7687 [8.4833, 9.0661]
Calibrated GT + Loop	8.7743 [8.4723, 9.0869]
Calibrated GT + Latent CoT	8.8254 [8.5293, 9.1307]
Calibrated GT + Staged Latent CoT	**8.5000** [8.1843, 8.8260]
Pooled CT	9.6089 [9.2232, 9.9987]
Pooled Mamba-CDSP	10.5912 [10.2028, 10.9831]

Table H.6: Terminal effect prediction for the CT transfer comparison. Entries are point estimates with pointwise 95% paired patient-bootstrap confidence intervals.

Model	**Effect RMSE,** $t = 16 \downarrow$
Pooled CT	9.6089 [9.2232, 9.9987]
Calibrated CT + Staged Latent CoT	**8.4254** [8.0863, 8.7720]

H.1 Dynamic-policy subgroup accuracy

This diagnostic evaluates whether the improvement from integrated UHC extends to subgroup-average effects under dynamic-policy evaluation, where actions are selected from generated histories rather than replayed. We reaggregate the saved dynamic-policy predictions of Pooled GT and Calibrated GT on the independent, quality-controlled test cohort of 5,997 patients. No additional training or model inference is performed.

We apply the fixed pretreatment routing rules learned from the training cohort to the test patients. The analysis includes all 26 retained calibration cells: two subgroups at each of 13 times, $t \in \{1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 14, 15, 16\}$. Neither the rules nor the retained cell set is reselected using test outcomes. Within each cell, both model predictions and simulator references are averaged over exactly the same UTG patients, matching the population used for the model-side training contrast. Test UCG outcomes are not used as the reference in this diagnostic.

Let $\mathcal{I}^{\text{test}}_{1,g,t}$ contain the retained UTG patients in cell (g, t), with size $n^{\text{test}}_{1,g,t}$. For model m, let $\widehat{\tau}^{m,\text{dyn}}_{i,t}$ denote its saved expected dynamic-policy effect relative to no-op, and let $\widetilde{\tau}_{i,t}$ denote the corresponding simulator MC reference. We compute

$$e^m_{g,t} = \frac{1}{n^{\text{test}}_{1,g,t}} \sum_{i \in \mathcal{I}^{\text{test}}_{1,g,t}} \left(\widehat{\tau}^{m,\text{dyn}}_{i,t} - \widetilde{\tau}_{i,t} \right), \qquad \text{RMSE}^m_{\text{group}} = \sqrt{\sum_{(g,t) \in \mathcal{C}_*} w_{g,t} (e^m_{g,t})^2},$$

where $\mathcal{C}_*$ is the fixed set of 26 cells. We give equal weight to each retained time and then to each cell within that time; here $w_{g,t} = 1/26$. This metric measures error in cell-average effects, not the individual-level Effect RMSE reported in the main comparison. Saved model estimates use 100 MC rollouts per policy; simulator references use eight per policy, as in the main evaluation.

The group-mean Effect RMSE decreases from **10.4021 mmHg** for Pooled GT to **2.3593 mmHg** for Calibrated GT, a **77.3%** reduction. Absolute error decreases in **all 26 cells** (Table H.7). These results establish an improvement in dynamic-policy subgroup prediction on this benchmark under integrated UHC. They do not compare replay and dynamic predictions directly or isolate the contribution of the contrast loss. The estimates are conditional on the fitted seed-2032 checkpoints and saved MC outputs; no additional confidence intervals are computed. Cells share patients across times and are not independent experimental replications.

Table H.7: Dynamic-policy subgroup effect prediction on the independent test cohort. The reference is the mean simulator MC effect over the same UTG patients used to average each model's predicted effect. Reference effects and absolute group-mean errors are in mmHg. All entries are point estimates from saved predictions. Cell identifiers enumerate the fixed training-derived subgroup–time cells; each retained time has two cells.

Cell	Time	n_{UTG}	MC reference effect	Pooled GT absolute error	Calibrated GT absolute error
1	1	444	9.8318	2.4958	0.6885
2	1	682	4.1038	0.9022	0.3788
3	2	724	10.8154	1.2845	0.8579
4	2	589	21.0058	1.0425	0.8945
5	3	450	28.0625	4.4132	2.4841
6	3	934	11.9748	2.6980	1.4001
7	4	199	31.0218	5.6407	2.8763
8	4	671	9.4875	3.6573	1.4193
9	5	644	12.5963	5.3698	2.3026
10	5	586	27.2964	7.2718	2.4395
11	6	699	11.6675	6.1425	2.3384
12	6	486	28.0665	10.5950	3.1271
13	9	522	19.8265	6.1969	1.8350
14	9	258	30.1819	15.5144	3.4191
15	10	434	27.4471	16.0546	2.8047
16	10	1049	12.6930	7.5931	2.5214
17	11	615	9.3151	8.7516	2.9026
18	11	396	26.2738	16.2923	2.8280
19	12	697	9.8334	8.0901	2.5083
20	12	439	25.5404	17.2401	2.8380
21	14	747	10.6059	7.6602	2.4264
22	14	569	23.4283	16.7521	2.4885
23	15	465	26.0682	18.6580	2.6437
24	15	861	10.6074	8.0701	2.3907
25	16	396	26.8494	19.5274	2.8288
26	16	960	11.3052	8.1710	2.1957

H.2 Anchor Calibration Ablation

This ablation evaluates the addition of subgroup state and outcome contrasts and within-arm distributional calibration to factual GT training. Anchor GT omits the individual-level free-rollout losses in both randomized arms while retaining generated trajectories for calibration. Anchor GT and Pooled GT share the GT architecture, initial weights, training patients, normalization-fitting patients, and residual-reference patients. Both use the same pooled factual validation loss for checkpoint selection; the selected checkpoints are epochs 80 and 42, respectively.

Evaluation protocol. This additional ablation uses training seed 2032 and the same 5,997 quality-controlled patients from the independent seed-2036 test cohort as the main comparison. Simulator references average eight realizations, and model predictions average 100 rollouts. Table H.8 places Anchor GT alongside Pooled GT, Calibrated GT (full UHC), and Calibrated GT + Staged Latent CoT. This ablation is reported separately from the six-comparator Bonferroni family in the main comparison.

Table H.8: Primary prediction metrics for the seed-2032 anchor-calibration ablation and the corresponding full-UHC and staged models. Entries are RMSEs in mmHg with pointwise 95% paired patient-bootstrap confidence intervals on the same 5,997 test patients. Lower is better; bold denotes the lowest point estimate.

Model	Effect RMSE, all times	Policy Y RMSE
Pooled GT	14.5826 [14.2731, 14.8974]	12.8461 [12.5666, 13.1323]
Anchor GT	10.3922 [10.0689, 10.7200]	10.4084 [10.0868, 10.7350]
Calibrated GT (full UHC)	10.1987 [9.9295, 10.4734]	10.0769 [9.8054, 10.3530]
Calibrated GT + Staged Latent CoT	**9.9358** [9.6477, 10.2283]	**9.8819** [9.5955, 10.1733]

Paired differences and interpretation. Relative to Pooled GT, Anchor GT yields paired differences of -4.1904 [-4.4974, -3.8860] in Effect RMSE and -2.4377 [-2.6595, -2.2182] in Policy Y RMSE. Brackets denote pointwise 95% confidence intervals from 100,000 paired patient-bootstrap resamples. These intervals are conditional on the fitted seed-2032 checkpoints and saved Monte Carlo estimates; they do not incorporate training-seed variation or additional Monte Carlo estimation uncertainty.

Both intervals exclude zero, supporting the benefit of combined contrast and distributional calibration without individual-level free-rollout supervision. Relative to Anchor GT, full UHC lowers the two primary RMSE point estimates by 1.86% and 3.18%, respectively; staged latent computation further lowers them by 2.58% and 1.94% relative to full UHC. The Anchor-to-full-UHC comparison reports point-estimate improvements; no paired difference interval is reported for that comparison. This ablation assesses contrast and distributional calibration jointly and does not identify their individual contributions. The matched staged-versus-unstaged comparison and its replication across training seeds are reported in Appendix I.

H.3 Supervision-Schedule Ablation

This ablation tests whether the assignment and order of intermediate targets improve counterfactual prediction beyond applying the same targets uniformly. All three conditions use the calibrated latent-CoT architecture with three passes, two feedback updates, and 1,215,827 parameters. They share initial weights, training patients, normalization-fitting patients, residual-reference patients, prediction heads, the final-pass UHC objective, and inference computation. Only the assignment of intermediate supervision changes.

Matched objectives. Write $L_X^{(r)} = \mathcal{L}_{X,\text{factual}}^{(r)}$ and $L_Y^{(r)} = \mathcal{L}_{Y,\text{factual}}^{(r)}$ for the observed-prefix losses defined in Appendix D. The three schedules are

$$\begin{aligned}
\mathcal{L}_{\text{Stage}} &= \mathcal{L}_{\text{UHC}}^{(2)} + 0.10L_X^{(0)} + 0.20L_Y^{(1)}, \\
\mathcal{L}_{\text{Uniform}} &= \mathcal{L}_{\text{UHC}}^{(2)} + 0.05\big(L_X^{(0)} + L_X^{(1)}\big) + 0.10\big(L_Y^{(0)} + L_Y^{(1)}\big), \\
\mathcal{L}_{\text{Reverse}} &= \mathcal{L}_{\text{UHC}}^{(2)} + 0.20L_Y^{(0)} + 0.10L_X^{(1)}.
\end{aligned}$$

Each schedule retains aggregate auxiliary coefficients of 0.10 for states and 0.20 for outcomes, including the same $1/17$ outcome-loss normalization. Uniform averages the losses across passes, not their predictions. The common final-pass objective includes factual prediction, free-rollout supervision, and subgroup contrast and distributional calibration.

Training and evaluation. All schedules use training seed 2032, an 80-epoch limit, patience 10, and batch size 512. The same pooled factual validation loss selects checkpoints: epoch 78 for Stage and epoch 80 for Uniform and Reverse. Counterfactual truth is used only for evaluation. We use the same 5,997 patients from the independent seed-2036 test cohort, freezing the reference-based exclusions from the original quality-controlled analysis (Appendix E). Simulator references average eight realizations and model predictions average 100 rollouts. The original Stage predictions are reused after checkpoint and evaluation-artifact verification.

Table H.9: Supervision-schedule ablation at training seed 2032. Entries are RMSEs in mmHg with pointwise 95% paired patient-bootstrap intervals on the same 5,997 test patients. All conditions use the same intermediate targets and aggregate loss weights. Bold denotes the lowest RMSE.

Supervision schedule	Effect RMSE, all times	Policy Y RMSE
Stage: state then outcome	**9.9358** [9.6477, 10.2283]	**9.8819** [9.5955, 10.1733]
Uniform: both targets at both passes	10.2336 [9.9517, 10.5207]	10.1111 [9.8291, 10.3996]
Reverse: outcome then state	10.2242 [9.9441, 10.5102]	10.1134 [9.8328, 10.4009]

Paired comparisons. We use 100,000 paired patient-bootstrap resamples, retaining each patient's complete trajectory and the same resample across models. Bonferroni correction covers all 18 evaluated metrics against both alternative schedules, forming a separate family of 36 comparisons. The intervals in Table H.10 are conditional on the fitted seed-2032 checkpoints, residual banks, and saved Monte Carlo estimates; they do not include training-seed variation or additional Monte Carlo estimation uncertainty.

Results and interpretation. State-then-outcome supervision reduces Effect RMSE and Policy Y RMSE by 2.91% and 2.27% relative to Uniform, and by 2.82% and 2.29% relative to Reverse. All four improvements remain significant after correction across 36 comparisons. Terminal Effect RMSE is also lower: 8.5000 for Stage, versus 8.8339 for Uniform and 8.8321 for Reverse. Uniform and Reverse achieve lower one-step state NRMSE (0.122698 and 0.122002) than Stage (0.124130), despite their higher counterfactual RMSEs. These results establish that the assignment and order of the same intermediate targets affect policy-trajectory and effect prediction, with the proposed state-then-outcome schedule outperforming both alternatives. This seed-2032 ablation complements the three-seed comparison against unstaged Latent CoT, which tests the addition of intermediate supervision (Appendix I).

Table H.10: Primary-metric differences for Stage minus each alternative schedule, in mmHg. Negative values favor Stage. Brackets contain Bonferroni-adjusted paired patient-bootstrap intervals with 95% familywise coverage across all 36 comparisons in this ablation. All four primary-metric intervals exclude zero.

Comparator	**Effect RMSE difference [adjusted CI]**	**Policy Y RMSE difference [adjusted CI]**
Uniform	−0.2979 [−0.4044, −0.1945]	−0.2292 [−0.3303, −0.1306]
Reverse	−0.2884 [−0.3933, −0.1875]	−0.2315 [−0.3317, −0.1347]

I Replication across Training Seeds

We compare Calibrated GT, Calibrated GT + Latent CoT, and Calibrated GT + Staged Latent CoT across training seeds 2031, 2032, and 2033. This comparison tests whether the gains from staged supervision reproduce across training runs. The unstaged and staged latent models share their architecture, parameter count, feedback updates, and inference computation; their only difference is intermediate state and outcome supervision during training.

Experimental protocol. We hold the source-training and validation cohorts, data partitions, constructed subgroups, calibration references, and evaluation references fixed across seeds. Each model retains the same hyperparameters and checkpoint-selection criterion, with only the training seed changed. Training follows Appendix F, including the 80-epoch limit and checkpoint selection by factual prediction loss on the observed validation cohort.

All runs are evaluated on the same 5,997 patients from the independent test cohort generated with seed 2036, using the same model-independent quality-control exclusions. Simulator reference trajectories average eight realizations, and model predictions average 100 rollouts. Effect RMSE and Policy Y RMSE across all prediction times remain the primary metrics; terminal Effect RMSE evaluates effect prediction at $T = 16$.

Training-seed variation and patient-sampling uncertainty. We summarize performance across the three training seeds using the arithmetic mean and sample standard deviation of the run-level RMSEs. We also calculate the difference between Staged Latent CoT and each comparator within each training seed. Patient-sampling uncertainty in these differences is estimated using 100,000 paired patient-bootstrap resamples, with complete patient trajectories resampled together and the same resample applied to each model.

For seeds 2031 and 2033, Bonferroni correction covers the two comparator models separately for each seed and metric. For seed 2032, we retain the intervals from Table H.4, which cover the six comparisons in the original main comparison separately for each primary metric. These intervals quantify patient-sampling uncertainty conditional on the fitted checkpoints and saved Monte Carlo estimates. Training-seed variation is reported separately through the across-seed standard deviations.

Table I.1: Performance across training seeds. All models use the calibrated GT backbone. "Unstaged Latent CoT" and "Staged Latent CoT" denote its corresponding extensions. Values are RMSEs in mmHg; lower is better. The final three rows report mean ± sample standard deviation across seeds. Bold denotes the lowest RMSE within each seed or the lowest across-seed mean.

Training seed	Model	Effect RMSE, all times	Policy Y RMSE	Terminal Effect RMSE
2031	Calibrated GT	10.1408	10.0458	8.6722
2031	Unstaged Latent CoT	10.1705	10.0436	8.7776
2031	Staged Latent CoT	**9.9388**	**9.8677**	**8.4333**
2032	Calibrated GT	10.1987	10.0769	8.7687
2032	Unstaged Latent CoT	10.2410	10.1399	8.8254
2032	Staged Latent CoT	**9.9358**	**9.8819**	**8.5000**
2033	Calibrated GT	10.0287	9.9922	8.5972
2033	Unstaged Latent CoT	10.3412	10.2501	8.8665
2033	Staged Latent CoT	**9.9921**	**9.9237**	**8.5527**
Mean ± SD	Calibrated GT	10.1227 ± 0.0864	10.0383 ± 0.0428	8.6794 ± 0.0860
Mean ± SD	Unstaged Latent CoT	10.2509 ± 0.0858	10.1445 ± 0.1033	8.8232 ± 0.0445
Mean ± SD	Staged Latent CoT	**9.9556 ± 0.0317**	**9.8911 ± 0.0291**	**8.4953 ± 0.0598**

Results across training seeds. Staged Latent CoT achieves the lowest Effect RMSE, Policy Y RMSE, and terminal Effect RMSE in all three training seeds (Table I.1). Comparing across-seed mean RMSEs, staged supervision reduces Effect RMSE by **2.88%** and Policy Y RMSE by **2.50%** relative to unstaged Latent CoT. Relative to Calibrated GT, the reductions are **1.65%** and **1.47%**, respectively. Mean terminal Effect RMSE decreases by **3.72%** relative to unstaged Latent CoT and **2.12%** relative to Calibrated GT.

Table I.2: Differences in the primary metrics within each training seed. Each difference is Staged Latent CoT minus the comparator; negative values favor Staged Latent CoT. Brackets contain Bonferroni-adjusted paired patient-bootstrap intervals with 95% familywise coverage within each seed and metric. Correction covers two comparisons for seeds 2031 and 2033 and the original six comparisons for seed 2032.

Training seed	Comparator	Effect RMSE difference [adjusted CI]	Policy Y RMSE difference [adjusted CI]
2031	Calibrated GT	−0.2020 [−0.2884, −0.1152]	−0.1781 [−0.2732, −0.0827]
2031	Unstaged Latent CoT	−0.2318 [−0.3329, −0.1327]	−0.1759 [−0.2826, −0.0719]
2032	Calibrated GT	−0.2629 [−0.3735, −0.1469]	−0.1950 [−0.3039, −0.0820]
2032	Unstaged Latent CoT	−0.3052 [−0.3923, −0.2212]	−0.2580 [−0.3418, −0.1768]
2033	Calibrated GT	−0.0365 [−0.1129, +0.0416]	−0.0685 [−0.1450, +0.0094]
2033	Unstaged Latent CoT	−0.3491 [−0.4555, −0.2437]	−0.3263 [−0.4298, −0.2236]

The advantage over unstaged Latent CoT is significant on both primary metrics in every training seed after Bonferroni correction (Table I.2). The corresponding paired differences have across-seed means and sample standard deviations of $\mathbf{-0.295 \pm 0.059}$ for Effect RMSE and $\mathbf{-0.253 \pm 0.075}$ for Policy Y RMSE.

Staged Latent CoT also improves both primary metrics over Calibrated GT in all three seeds. These differences are significant after correction for seeds 2031 and 2032; for seed 2033, both adjusted intervals include zero. The across-seed paired differences are $\mathbf{-0.167 \pm 0.117}$ for Effect RMSE and $\mathbf{-0.147 \pm 0.069}$ for Policy Y RMSE.

The comparison with unstaged Latent CoT isolates the contribution of intermediate supervision within the same recurrent architecture. Its improvement on both primary metrics across all three training seeds shows that staged supervision yields reproducible accuracy gains without increasing parameter count or inference computation.

J Generalization to an Unseen Policy

Experimental protocol. Models were trained under a MAP-target-90 policy. We evaluated the fixed seed-2032 checkpoints of Pooled GT and Staged Latent CoT with UHC under an unseen MAP-target-80 policy, without retraining, recalibration, or checkpoint reselection. Both models were evaluated on the same 5,997 quality-controlled test patients used in the main comparison. No patients were excluded based on outcomes or model errors under the alternative policy. At each time, both models selected actions from their own recursively generated histories, without future observed trajectories. Simulator references average eight realizations of future innovations; model predictions average 100 rollouts. Under the target-80 policy, the reported action proportions are 0.9275 no-op, 0.0354 fluid, and 0.0371 vasopressor, and 57.29% of patients receive at least one intervention.

The prespecified unseen-policy evaluation comprised only the MAP-target-80 policy. Model-specific intervals and intervals for paired differences are pointwise 95% paired patient-bootstrap intervals,

reported separately for each metric. Each metric involves one comparison between the two models under this single prespecified policy, so no multiplicity correction across policies is applied. These intervals are conditional on the fixed seed-2032 checkpoints and saved MC estimates; they do not include training-seed variation or additional MC estimation uncertainty.

Table J.1: Prediction under the unseen MAP-target-80 policy. RMSEs are in mmHg; brackets contain pointwise 95% paired patient-bootstrap intervals. Bold denotes the lower point estimate. Relative reductions compare the complete method with Pooled GT.

Metric	Pooled GT	Staged Latent CoT with UHC	Relative reduction
Effect RMSE, all times	11.0543 [10.7708, 11.3448]	**9.5740** [9.2954, 9.8554]	13.39%
Effect RMSE, terminal	12.3330 [12.0351, 12.6344]	**8.2191** [7.9810, 8.4618]	33.36%
Policy Y RMSE	9.8294 [9.5716, 10.0950]	**9.5719** [9.2873, 9.8598]	2.62%

Table J.2: Paired RMSE differences under the prespecified unseen MAP-target-80 policy. Differences are Staged Latent CoT with UHC minus Pooled GT; negative values favor the complete method. Intervals are pointwise 95% paired patient-bootstrap confidence intervals for each metric.

Metric	Difference	Paired 95% CI
Effect RMSE, all times	−1.4803	[−1.7559, −1.2040]
Effect RMSE, terminal	−4.1138	[−4.3926, −3.8349]
Policy Y RMSE	−0.2575	[−0.5241, +0.0109]

Results. The complete method reduces all-time Effect RMSE by **13.39%** and terminal Effect RMSE by **33.36%** (Table J.1). Both improvements are significant, with paired 95% confidence intervals excluding zero (Table J.2). Policy Y RMSE decreases by 2.62%, although its paired 95% difference interval includes zero. These results demonstrate transfer of the complete method's effect-prediction gains to the evaluated unseen policy. This comparison evaluates Staged Latent CoT with UHC as a whole; it does not isolate staged supervision or establish identification of arbitrary untested policy effects.