VASODPO: ALIGNING TREATMENT ACTIONS WITH PATIENT STATES THROUGH STATE-FIXED PREFERENCE CONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical treatment generation aims to tailor actions to patients, yet retrospective electronic health records provide limited preference supervision over alternative actions under the same patient state. We study this problem through retrospective single-step generation of normalized norepinephrine actions for patients with sepsis in intensive care units. We propose VasoDPO, a state-fixed Direct Preference Optimization (DPO) framework that anchors chosen responses to recorded actions and uses mean arterial pressure (MAP) to construct rejected actions under the same patient state. After minimum action-gap filtering, VasoDPO upweights retained pairs contrasting a nonzero recorded action with a zero rejected action, motivated by near-zero concentration observed after supervised fine-tuning. Because favorable action-replacement train-on-synthetic, test-on-real (TSTR) scores can coexist with near-zero action concentration when state features are retained, we jointly evaluate TSTR with direct diagnostics of near-zero concentration, directional violations, and MAP sensitivity. Across five training seeds on a fixed split, VasoDPO achieves higher mean TSTR AUROC and AUPRC and lower mean near-zero and directional-violation rates than a matched DPO baseline using the same preference pairs with uniform weights. Matched ablations support the contributions of MAP-guided construction and omission weighting, while state–pair shuffling degrades both TSTR and action diagnostics. Controlled MAP sweeps show that the mean generated action varies with MAP while other inputs remain fixed.

1 INTRODUCTION

Clinical treatment generation produces patient-specific actions from heterogeneous clinical data, including physiological measurements, laboratory results, and treatment histories recorded in electronic health records (EHRs) (Ye et al., 2024). Applications include medication recommendation, chronic disease management, and critical care (Tan et al., 2024; Li et al., 2024; Komorowski et al., 2018). In critical care, rapid physiological changes make the patient's clinical state central to treatment selection (Raghu et al., 2017; Freyer et al., 2026).

A core research problem in this domain is how to generate treatment actions tailored to individual patient states (Kim et al., 2025; Deng et al., 2026). Patients with similar diagnoses or clinical pathways can differ in their physiology and responses to treatment (Niraula et al., 2025; Matsson et al., 2025). This heterogeneity motivates treatment generators that respond to clinically relevant differences in patient state. We study this problem in a retrospective single-step setting, where each action is conditioned on a recorded patient state.

However, learning such state-sensitive behavior from retrospective records is challenging because supervision for comparing alternative actions under a given state is limited. Recorded clinician actions support supervised fine-tuning (SFT) of large language models (LLMs) to predict observed actions from patient states. To complement this supervision, pairwise preference learning introduces explicit comparisons between candidate actions under the same patient state. Direct Preference Optimization (DPO) provides an objective for optimizing a policy from such chosen and rejected response pairs (Rafailov et al., 2023). Applying DPO to retrospective clinical records raises two practical challenges. First, DPO operates on chosen and rejected response pairs, whereas retrospective

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.

records typically provide only the administered action for each recorded state. Therefore, a state-matched alternative must be constructed and assigned to the rejected side of the comparison before preference optimization. Second, a binary preference specifies the ordering within each pair but leaves the relative training emphasis across different action contrasts unspecified. Therefore, the training design must determine how different comparisons are prioritized during optimization.

We propose VasoDPO, a state-fixed preference optimization framework with rejected-action construction and pair weighting. We apply VasoDPO to normalized norepinephrine action generation for patients with sepsis in intensive care units (ICUs), where vasopressor treatment depends on the patient's hemodynamic state (Raghu et al., 2017; Komorowski et al., 2018; Evans et al., 2021). For each recorded patient state, VasoDPO anchors the chosen response to the recorded action and constructs a MAP-guided rejected action under the same state. We retain pairs meeting a prespecified minimum action gap. Motivated by near-zero concentration in SFT, we upweight retained pairs with a nonzero recorded action and a zero rejected action. The resulting pair weights are normalized to unit mean over the preference corpus and used in the logistic DPO objective.

We evaluate generated actions using downstream predictive performance and direct action diagnostics. For downstream prediction, we use an action-replacement train-on-synthetic, test-on-real (TSTR) protocol adapted from synthetic-data evaluation (Kotelnikov et al., 2023; Lautrup et al., 2025; Niu et al., 2025). We replace recorded training actions with generated actions while retaining states and outcomes, train an in-hospital mortality predictor, and evaluate it on held-out real records using AUROC and AUPRC. Because the downstream predictor retains the original state features, favorable TSTR scores can coexist with near-zero concentration in the generated actions. To complement TSTR, we assess generated actions directly on held-out states: the near-zero rate (NZR) measures the fraction of valid generated actions near zero. The critical generation violation rate (CGVR) aggregates violations defined by thresholds on standardized MAP and normalized action values. A controlled MAP audit measures the mean action response to MAP while holding all other inputs fixed.

Our contributions are summarized as follows:

- We propose VasoDPO, a state-fixed preference optimization framework for retrospective treatment generation. It constructs rejected actions using MAP under the recorded patient state, filters pairs by a minimum action gap, and applies corpus-normalized weights to emphasize nonzero recorded actions paired with zero alternatives.
- We show that favorable action replacement TSTR scores can coexist with near-zero action concentration. Constant action controls obtain favorable scores on selected metrics, while shuffled actions retain diversity despite disrupted correspondence with patient states.
- Across five training seeds on a fixed ICU sepsis split, VasoDPO achieves higher mean TSTR AUROC and AUPRC and lower mean NZR and CGVR than DPO trained on the same preference pairs with uniform weights. Matched ablations support the contributions of MAP guidance and omission weighting. Shuffling patient states across preference pairs worsens both predictive and action diagnostic scores.

2 Related Work

Clinical treatment modeling. Medication recommendation and clinical policy models use patient health status and history to inform treatment predictions (Kim et al., 2025; Deng et al., 2026; Matsson et al., 2025). In sepsis treatment planning, reinforcement learning methods model treatment as a sequence of decisions and evaluate policies using learned dynamics, simulation, or off-policy evaluation (Raghu et al., 2017; Komorowski et al., 2018; Li et al., 2019; Wu et al., 2026; Gottesman et al., 2019). Recent medical studies apply DPO to radiology report generation and benchmark its variants for medical vision–language models, while a clinical LLM study compares DPO with SFT (Banerjee et al., 2024; Kim et al., 2026a; Savage et al., 2025). We study retrospective single-step generation, with a recorded patient state as input and a normalized treatment dose as output.

Synthetic-data evaluation. Methods for synthesizing EHRs and tabular data generate structured records (Choi et al., 2017; Xu et al., 2019; Kotelnikov et al., 2023). Evaluation frameworks assess distributional fidelity and downstream predictive utility (Lautrup et al., 2025; Niu et al., 2025; Stoian et al., 2024). For downstream causal inference, Amad et al. (2025) assess whether synthetic

data preserve the covariate distribution, treatment assignment mechanism, and outcome generation mechanism. Our action-replacement evaluation retains patient states and outcomes while replacing recorded actions with generated doses. We combine downstream mortality prediction with direct action diagnostics and a controlled MAP audit.

Preference optimization. DPO learns from chosen and rejected responses conditioned on a shared input using a logistic preference objective (Rafailov et al., 2023). SimPO uses a reference-free reward formulation (Meng et al., 2024), and KTO learns from binary desirability labels (Ethayarajh et al., 2024). SafeDPO incorporates safety constraints into preference optimization (Kim et al., 2026b). Alignment-Weighted DPO assigns separate preference weights to reasoning and final-answer segments (Hu et al., 2026), while TI-DPO uses attribution-guided token weights and a triplet objective (Yang et al., 2026b). VasoDPO weights complete action comparisons, emphasizing pairs with nonzero recorded actions and zero rejected actions.

Preference construction. ACT uses simulated conversations to construct contrasts between dialogue actions in a quasi-online preference learning procedure (Chen et al., 2025). BMC constructs a pseudo-winning response by improving a losing response with the winning response as a reference, and models token-level correlations within preference pairs (Jiang et al., 2025). SSPO supplements a small set of labeled preferences with unpaired samples through pseudo-labeling (Lee et al., 2026). MetaAPO jointly adapts online data generation and sample weighting (Yang et al., 2026a), while corrective intervention constructs preference pairs by editing safety-relevant reasoning steps (Zhang et al., 2026). For continuous control, PREFINE constructs policy-sampled comparison trajectories to fine-tune policies from trajectory-level preferences (Verma et al., 2026). VasoDPO constructs alternatives in the normalized action space, using recorded actions as chosen anchors and MAP to guide rejected candidates.

3 METHOD

VasoDPO trains a treatment generator using preference pairs constructed under fixed recorded patient states and weighted within the DPO objective. For each eligible record, a MAP-guided operator pairs the recorded action with a constructed alternative under the same serialized state. We retain pairs whose numeric actions differ by at least a minimum action gap and upweight retained comparisons between a nonzero recorded action and a zero rejected action. Figure 1 summarizes the workflow, and Table 3 in Appendix A lists the notation.

3.1 PROBLEM SETUP

Let $\mathcal{D}_{\text{train}} = \{(x_i, a_i, y_i, g_i)\}_{i=1}^{N}$ denote the N training records. Here, x_i is the clinical context, $a_i \in [0, 1]$ is the observed normalized norepinephrine action, $y_i \in \{0, 1\}$ is the in-hospital mortality outcome, and g_i denotes the released trajectory group. We use a_i as the generation target, y_i as the mortality target for downstream TSTR evaluation, and g_i to construct group-disjoint data splits.

The field selector Ψ retains the model-visible clinical inputs, and the serializer Φ formats them with a fixed task instruction into the patient state S_i. The autoregressive policy $\pi_\theta(\cdot \mid S_i)$, with trainable parameters θ, defines a distribution over response sequences. For the main LLM comparison, greedy decoding produces the response r_i. State construction and response parsing are written as

$$S_i = \Phi(\Psi(x_i)), \qquad (q_i, \hat{a}_i) = \text{ParseAction}(r_i). \tag{1}$$

The parser (Appendix A.2) returns $q_i \in \{0, 1\}$, indicating whether parsing succeeds. For $q_i = 1$, the extracted numeric value is clipped to $[0, 1]$ to obtain $\hat{a}_i$.

3.2 STATE REPRESENTATION AND SERIALIZATION

The serialized state contains four released fields: the released APACHE-derived severity covariate (`apache`), scaled time, standardized heart rate, and standardized mean arterial pressure (MAP). The APACHE score is constant within each trajectory group, while the other fields describe individual records. Numeric inputs in S_i are formatted to three decimal places, and missing values and APACHE sentinel values are represented using fixed text labels. All learned LLM variants share Ψ and Φ; Appendix A.1 specifies the field order, labels, and formatting rules.

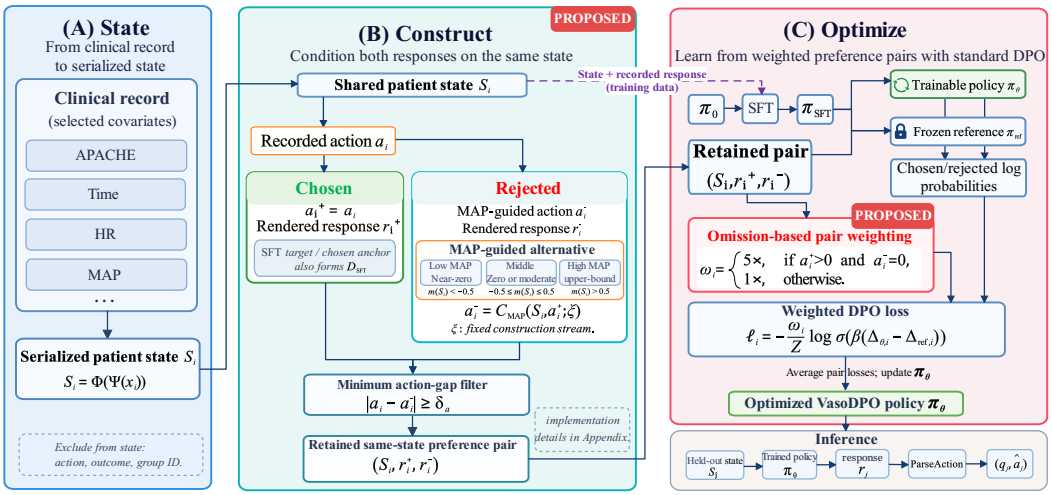

Figure 1: **VasoDPO overview.** (A) Clinical covariates form the shared state. (B) MAP-guided alternatives are paired with recorded actions and filtered by action gap. (C) SFT initializes both policies; corpus-normalized omission weights scale the DPO loss, with the reference policy frozen.

3.3 State-Fixed, Action-Local Preference Construction

Let $\mathcal{I}_{\mathrm{SFT}} \subseteq \{1, \ldots, N\}$ index records with both an observed action and an observed MAP value. For each eligible record, we set $a_i^+ = a_i$. The deterministic function $\mathrm{RenderAction}(a)$ formats a numeric action to three decimal places within a fixed response template. The rendered recorded action, $r_i^+ = \mathrm{RenderAction}(a_i^+)$, serves as both the SFT target and the chosen response in preference training. The corresponding supervised corpus is $\mathcal{D}_{\mathrm{SFT}} = \{(S_i, r_i^+)\}_{i \in \mathcal{I}_{\mathrm{SFT}}}$. Minimizing the completion-only negative log-likelihood on this corpus yields the SFT policy π_{SFT}. Starting from the same eligible records, we construct a candidate rejected action a_i^- while keeping S_i fixed.

Let $m(S_i)$ denote the standardized MAP value in the underlying record before three-decimal serialization. Thresholds $\tau_{\mathrm{MAP}}^- = -0.5$ and $\tau_{\mathrm{MAP}}^+ = 0.5$ define the low- and high-MAP indicators:

$$L_i = \mathbf{1}\{m(S_i) < \tau_{\mathrm{MAP}}^-\}, \qquad H_i = \mathbf{1}\{m(S_i) > \tau_{\mathrm{MAP}}^+\}. \tag{2}$$

Here $\mathbf{1}\{\cdot\}$ is the indicator function. States satisfying $L_i = H_i = 0$, including both threshold boundaries, form the intermediate group.

The operator C_{MAP} constructs a candidate rejected action from the MAP group and recorded action using the following rules. The low- and high-MAP branches generate alternatives near the lower and upper action bounds, respectively, while the intermediate branch depends on the recorded action. In the low-MAP group, it selects zero with probability 0.7 and otherwise samples uniformly from $[0.001, 0.020]$. In the high-MAP group, it selects one with probability 0.7 and otherwise samples uniformly from $[0.70, 0.99]$. In the intermediate group, it samples uniformly from $[0.20, 0.60]$ when $a_i^+ \le 0.05$, and selects zero otherwise. Using a pseudorandom stream ξ initialized with seed 42, we construct the rejected action and render both responses with the same template:

$$\begin{aligned} a_i^- &= C_{\mathrm{MAP}}(S_i, a_i^+; \xi), \\ r_i^+ = \mathrm{RenderAction}(a_i^+), &\qquad r_i^- = \mathrm{RenderAction}(a_i^-). \end{aligned} \tag{3}$$

Here r_i^- is the rendered rejected response. We assign the training preference $r_i^+ \succ_{S_i} r_i^-$, with the recorded action as chosen and the constructed alternative as rejected. Both responses share the state S_i and the response template; only the rendered action value changes. Appendix A.2 specifies the template. We retain pairs whose numeric actions differ by at least δ_a:

$$\mathcal{I}_{\mathrm{pref}} = \{i \in \mathcal{I}_{\mathrm{SFT}} : |a_i^+ - a_i^-| \ge \delta_a\}, \qquad \delta_a = 0.1, \tag{4}$$

The gap is computed from numeric actions before rounding them to the response format. Each eligible record contributes one candidate pair, which is retained only if its action gap meets δ_a.

MAP predicates and evaluation. The MAP-norepinephrine relation motivates using MAP to construct directional action contrasts: norepinephrine is the recommended first-line vasopressor for adults with septic shock, for whom MAP is a central treatment target (Evans et al., 2021). We define τ^{-}_{MAP} and τ^{+}_{MAP} as standardized protocol coordinates for preference construction and diagnostic eligibility. Section 4 defines the corresponding action thresholds and rates.

3.4 Omission-Based Pair Prioritization

To address the concentration near zero observed after SFT, we give greater weight to retained pairs that compare a nonzero recorded action with a zero rejected action. For each $i \in \mathcal{I}_{\mathrm{pref}}$, we define the omission indicator h_i^{omit}, the raw pair weight ω_i, and the corpus mean weight Z:

$$\begin{aligned} h_i^{\mathrm{omit}} &= \mathbf{1}\{a_i^{+} > 0 \wedge a_i^{-} = 0\}, \\ \omega_i = \omega_{\mathrm{omit}} h_i^{\mathrm{omit}} + \omega_{\mathrm{std}}(1 - h_i^{\mathrm{omit}}), &\qquad \omega_{\mathrm{omit}} > \omega_{\mathrm{std}} > 0, \\ Z &= \frac{1}{|\mathcal{I}_{\mathrm{pref}}|} \sum_{i \in \mathcal{I}_{\mathrm{pref}}} \omega_i. \end{aligned} \tag{5}$$

Here $h_i^{\mathrm{omit}} \in \{0, 1\}$ identifies whether pair i is an omission contrast. The raw weight ω_i is ω_{omit} for omission pairs and ω_{std} for all other retained pairs. When $a_i^{-} = 0$, the gap criterion in Eq. 4 implies $a_i^{+} \geq \delta_a$. The indicator is computed from numeric actions before rendering and requires $a_i^{-} = 0$ exactly. It applies to retained pairs from all MAP groups. Dividing by Z gives normalized weights ω_i / Z with mean one over the retained pairs. The resulting weighted preference corpus is

$$\mathcal{D}_{\mathrm{pref}} = \{(S_i, r_i^{+}, r_i^{-}, \omega_i)\}_{i \in \mathcal{I}_{\mathrm{pref}}}. \tag{6}$$

This corpus is constructed once and reused across training seeds. We initialize π_θ from π_{SFT} and fix $\pi_{\mathrm{ref}} = \pi_{\mathrm{SFT}}$. For each retained pair, the chosen-versus-rejected log-probability differences are

$$\begin{aligned} \Delta_{\theta,i} &= \log \pi_\theta(r_i^{+} \mid S_i) - \log \pi_\theta(r_i^{-} \mid S_i), \\ \Delta_{\mathrm{ref},i} &= \log \pi_{\mathrm{ref}}(r_i^{+} \mid S_i) - \log \pi_{\mathrm{ref}}(r_i^{-} \mid S_i). \end{aligned} \tag{7}$$

Here $\log \pi(r \mid S)$ sums next-token log probabilities over response tokens, including the shared response terminator, with prompt tokens masked. The difference $\Delta_{\theta,i} - \Delta_{\mathrm{ref},i}$ measures the change in the chosen-to-rejected log probability ratio relative to the frozen SFT reference. With the logistic sigmoid $\sigma(z) = (1 + e^{-z})^{-1}$ and DPO scale $\beta > 0$, we minimize

$$\mathcal{L}_{\mathrm{VasoDPO}}(\theta) = -\frac{1}{|\mathcal{I}_{\mathrm{pref}}|} \sum_{i \in \mathcal{I}_{\mathrm{pref}}} \frac{\omega_i}{Z} \log \sigma(\beta[\Delta_{\theta,i} - \Delta_{\mathrm{ref},i}]). \tag{8}$$

Minimizing this objective encourages the policy to increase the chosen-to-rejected likelihood ratio relative to the frozen SFT reference, with greater emphasis on pairs receiving larger normalized weights. We use $\beta = 0.1$ and compute Z once over the complete preference corpus. During minibatch optimization, we update θ while keeping π_{ref} and the pair weights fixed. Appendix B.2 gives the gradient.

3.5 Training Procedure

We use `Meta-Llama-3-8B-Instruct` as the base policy π_0 (Grattafiori et al., 2024). Both SFT and preference optimization use low-rank adaptation (LoRA) (Hu et al., 2022); Appendix D reports the adapter configurations. For each training seed $k \in \mathcal{K} = \{42, 43, 44, 45, 46\}$, SFT trains on the same 117,812 eligible records with learning rate 2×10^{-5}, yielding $\pi_{\mathrm{SFT}}^{(k)}$. The seed index is omitted from the preceding equations for readability. VasoDPO uses this checkpoint for both initialization and the frozen reference policy. Matched baseline initialization is specified in Section 4. Using construction seed 42, we retain 76,442 preference pairs and keep them fixed across training seeds.

VasoDPO trains for three epochs with $\beta = 0.1$, learning rate 2×10^{-6}, and $(\omega_{\mathrm{omit}}, \omega_{\mathrm{std}}) = (5, 1)$. Algorithm 1 summarizes training, and Appendix D.5 reports the complete optimization settings.

4 EXPERIMENTAL PROTOCOL

Dataset and cohort. We use the ICU Sepsis component of the PhysioNet Clinical Time Series Datasets for Trajectory Flow Matching Evaluation, version 1.0.0 (Pu et al., 2026; Zhang et al., 2024; Pollard et al., 2026). This is a processed cohort derived from the eICU Collaborative Research Database (Pollard et al., 2018). We retain action-associated rows from the first 24 hours after ICU admission. Each record contains scaled time, standardized heart rate and MAP, and an APACHE-derived severity covariate (`apache`) shared by all records within its trajectory group. The target action is the normalized norepinephrine value (`norepi_inf_scaled`$\in [0, 1]$), and the outcome is in-hospital mortality (`HOSP_MORT`).

We pool all records from the released training, validation, and test partitions and construct a fixed split for retrospective action generation. We stratify trajectory groups by in-hospital mortality and assign each released `HADM_ID` group entirely to one partition using seed 42, targeting an approximately 80/20 ratio of training to test records. This yields 117,812 training records from 2,698 groups and 29,400 test records from 664 groups. Training records supply the SFT and preference corpora. Test records provide the real evaluation data for TSTR and the patient states for generated-action diagnostics. Appendix C.1 details the partition construction and preprocessing.

Downstream predictive evaluation. For each training seed, the fitted action model generates actions for the training partition. In action-replacement TSTR, these actions replace the recorded norepinephrine values while states and mortality outcomes remain fixed. An XGBoost mortality predictor (Chen & Guestrin, 2016) uses the four clinical covariates and the action value to predict in-hospital mortality. We evaluate AUROC and AUPRC on held-out real records containing recorded actions. The train-on-real, test-on-real (TRTR) reference uses the original training records and the same test records. A state-only predictor uses the four covariates to assess the contribution of the action feature. All predictors share the evaluator configuration in Appendix C.2. Held-out generated actions are evaluated separately using the diagnostics below. For learned action models, we compute each metric for the five training seeds and report its mean and standard deviation on the fixed split.

Primary diagnostics and normalized predicates. Low-MAP withholding (LMW) is the fraction of low-MAP records with an action near zero or a parse failure. High-MAP over-escalation (HME) is the fraction of high-MAP records with an action near the upper bound or a parse failure. CGVR aggregates violations across both groups, while NZR is the fraction of successfully parsed actions near zero. All held-out records have observed MAP values. For M held-out records indexed by $j = 1, \ldots, M$, let $q_j \in \{0, 1\}$ indicate parser acceptance and let $\hat{a}_j$ denote the parsed action when $q_j = 1$. Equation 2 defines the low- and high-MAP indicators L_j and H_j. We define the parser-valid and MAP-eligible index sets as follows:

$$\mathcal{V} = \{j : q_j = 1\}, \quad \mathcal{E}_{\text{low}} = \{j : L_j = 1\}, \quad \mathcal{E}_{\text{high}} = \{j : H_j = 1\}, \quad \mathcal{E}_{\text{crit}} = \mathcal{E}_{\text{low}} \cup \mathcal{E}_{\text{high}}. \tag{9}$$

Intermediate-MAP records contribute to NZR when parsing succeeds and are excluded from the directional eligibility sets. We use MAP thresholds $\tau^-_{\text{MAP}} = -0.5$ and $\tau^+_{\text{MAP}} = 0.5$, and normalized action thresholds $\tau^0_a = 0.01$ and $\tau^1_a = 0.99$. The threshold τ^0_a identifies generated actions near zero for evaluation. During training, omission weighting applies to retained pairs with $a^+_i > 0$ and $a^-_i = 0$, as defined in Eq. 5. For records in the low and high MAP groups, respectively, define the violation indicators:

$$d^{\text{low}}_j = \mathbf{1}\{q_j = 0 \vee (q_j = 1 \wedge \hat{a}_j \le \tau^0_a)\}, \qquad d^{\text{high}}_j = \mathbf{1}\{q_j = 0 \vee (q_j = 1 \wedge \hat{a}_j \ge \tau^1_a)\}. \tag{10}$$

Within each MAP eligibility group, a parse failure counts as a violation without evaluating a numeric action comparison. The four diagnostic rates are defined as follows:

$$\begin{aligned} \text{LMW} &= \frac{\sum_{j \in \mathcal{E}_{\text{low}}} d^{\text{low}}_j}{|\mathcal{E}_{\text{low}}|}, & \text{HME} &= \frac{\sum_{j \in \mathcal{E}_{\text{high}}} d^{\text{high}}_j}{|\mathcal{E}_{\text{high}}|}, \\ \text{CGVR} &= \frac{\sum_{j \in \mathcal{E}_{\text{low}}} d^{\text{low}}_j + \sum_{j \in \mathcal{E}_{\text{high}}} d^{\text{high}}_j}{|\mathcal{E}_{\text{crit}}|}, & \text{NZR} &= \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \mathbf{1}\{\hat{a}_j \le \tau^0_a\}. \end{aligned} \tag{11}$$

CGVR combines LMW and HME in proportion to the sizes of their eligibility groups. We report all four rates as percentages. Parser acceptance is $|\mathcal{V}|/M$ and equals 100% for all methods in the main comparison. Appendix C.3 specifies empty-set handling and the evaluator implementation; Appendix C.5 describes threshold sensitivity and high-MAP support audits.

Matched training variants and baselines. We compare two rejected-action constructions (generic and MAP-guided) under two weighting schemes (uniform and omission-based); Table 2 reports the results. Generic-action DPO-U and Generic-action DPO share the generic preference pairs and use uniform and omission weights, respectively. Standard-DPO uses uniform weights on the MAP-guided pairs, whereas VasoDPO applies omission weights to those same pairs. These four configurations use the same SFT initialization for each training seed and common optimization and decoding settings. This design assesses MAP guidance under each weighting scheme and omission weighting under each construction scheme. Appendix D.3 states their configuration mapping.

Context-shuffled DPO tests state–pair alignment by reassigning states to VasoDPO's response pairs while keeping each chosen response, rejected response, and pair weight together. Weighted SFT continues supervised training on the chosen responses from VasoDPO's retained corpus using the same omission weights. Additional comparisons include SFT-only, SimPO, KTO, SafeDPO-CGVR, Rule-rerank SFT, and the tabular action predictors Ridge (Hoerl & Kennard, 1970), XGBoost, and TabICLv2 (Qu et al., 2026). All methods share the downstream evaluator configuration and action-diagnostic definitions. Appendix D gives the training and inference settings. Separate diagnostic controls use constant zero, the training median, a MAP-dependent median rule, or shuffled training actions (Appendix E.1).

Controlled MAP audit. We assess MAP sensitivity using 100 held-out patient states from distinct trajectory groups. These anchor states are selected independently of model outputs and shared across methods and training seeds. For each anchor, we vary standardized MAP over nine values from -2 to 2 in increments of 0.5, holding all other input fields and decoding settings fixed. At each MAP value m, $\bar{a}(m)$ denotes the mean generated action across the 100 anchors and five training seeds. We report the mean response curve and the signed endpoint difference $\Delta_{\mathrm{MAP}} = \bar{a}(2) - \bar{a}(-2)$, which summarizes the direction and magnitude of the change in mean action between the endpoints. To quantify uncertainty across anchors and training seeds on the fixed split, we compute 95% confidence intervals from 2,000 bootstrap replicates, resampling both sets of indices with replacement and using the same draws across methods. Figure 2 presents the response curves. Appendix C.4 details anchor selection, bootstrap estimation, and MAP coverage in the held-out cohort.

5 Results

Main comparison. Table 1 shows that favorable TSTR performance can coexist with concentrated action generation. SFT-only achieves mean AUROC/AUPRC of $0.674/0.295$, close to the TRTR reference of $0.675/0.296$, while its mean NZR reaches 97.330% and its mean LMW reaches 97.589%. This contrast motivates evaluating the generated actions alongside downstream mortality prediction.

Comparing VasoDPO with Standard-DPO on the same preference pairs assesses the contribution of omission weighting. Across five training seeds, mean TSTR AUROC/AUPRC increases from $0.678/0.289$ to $0.697/0.311$, while mean CGVR decreases from 9.286% to 0.510% and mean NZR from 13.295% to 0.967%. Among the other LLM baselines, KTO achieves the highest mean TSTR AUROC/AUPRC at $0.686/0.300$, with a mean NZR of 43.043%. SafeDPO-CGVR achieves the lowest mean CGVR/NZR among these baselines, at 0.874% and 4.824%, respectively.

Ridge achieves zero CGVR and NZR, with 85.582% of its outputs in the predefined medium action bin. The standard deviation of its generated actions is 0.0172, compared with 0.0452 for VasoDPO (Appendix Tables 6 and 9). Relative to the state-only predictor, adding generated actions under the replacement protocol lowers mean AUROC/AUPRC by $0.022/0.029$ for Ridge and raises them by $0.022/0.017$ for VasoDPO (Appendix Table 10).

At the primary action threshold of 0.99, only 3 of the 7,462 recorded actions in high-MAP test states reach the threshold. None of the evaluated generated actions in these states reaches it (Appendix Table 8, Panel B). With HME equal to zero, differences in CGVR reflect differences in LMW, scaled by the proportion of low-MAP states among all low- and high-MAP eligible states.

5.1 Training comparisons and diagnostic interpretation

Table 2 compares rejected-action construction and pair weighting in Panel A. Panel B examines state assignment to preference pairs and compares preference training with Weighted SFT.

Table 1: **Main ICU Sepsis comparison.** Five-seed means ± standard deviations. Bold/underline: best/second-best mean among learned LLMs. Full generation audit: Appendix Table 6.

Method	AUROC ↑	AUPRC ↑	CGVR (%) ↓	NZR (%) ↓	LMW (%) ↓
Real-data reference					
TRTR reference	0.675	0.296	–	–	–
Tabular action policies					
Ridge	0.653 ± 0.000	0.264 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
XGBoost	0.672 ± 0.002	0.284 ± 0.002	1.586 ± 0.046	1.306 ± 0.022	2.908 ± 0.084
TabICLv2	0.667 ± 0.002	0.276 ± 0.002	2.358 ± 0.042	2.200 ± 0.034	4.325 ± 0.077
LLM action policies					
SFT-only LLM	0.674 ± 0.008	0.295 ± 0.005	53.216 ± 1.210	97.330 ± 2.509	97.589 ± 2.219
Weighted SFT	0.677 ± 0.003	0.294 ± 0.005	1.010 ± 0.904	68.369 ± 0.653	1.852 ± 1.658
Standard-DPO	0.678 ± 0.004	0.289 ± 0.003	9.286 ± 4.009	13.295 ± 5.072	17.030 ± 7.352
Rule-rerank SFT	0.676 ± 0.002	0.292 ± 0.002	34.771 ± 3.716	75.846 ± 6.579	63.764 ± 6.814
SimPO	0.674 ± 0.007	0.294 ± 0.004	45.773 ± 2.825	75.847 ± 2.964	83.940 ± 5.180
KTO	$\underline{0.686} \pm 0.004$	$\underline{0.300} \pm 0.003$	1.409 ± 1.001	43.043 ± 4.990	2.583 ± 1.836
SafeDPO-CGVR	0.678 ± 0.002	0.289 ± 0.004	$\underline{0.874} \pm 0.722$	$\underline{4.824} \pm 3.394$	$\underline{1.602} \pm 1.325$
VasoDPO	$\mathbf{0.697} \pm 0.005$	$\mathbf{0.311} \pm 0.006$	$\mathbf{0.510} \pm 0.033$	$\mathbf{0.967} \pm 0.013$	$\mathbf{0.935} \pm 0.061$

NZR uses parsed actions; LMW uses low-MAP eligible states; CGVR uses all low/high-MAP eligible states. TRTR uses real training and test data and has no generated-action diagnostics.

Table 2: **Factor ablations and matched controls.** (A) MAP guidance crossed with omission weighting. (B) State assignment and weighted supervised continuation. Five-seed means ± standard deviations under the fixed protocol.

A. Core 2×2 *factor comparison*

	Design factors				TSTR association		Action diagnostics	
Variant	Align.	MAP	Omit.	Pair.	AUROC ↑	AUPRC ↑	CGVR (%) ↓	NZR (%) ↓
Generic-action DPO-U	✓	–	–	✓	0.667 ± 0.008	0.279 ± 0.006	15.831 ± 4.832	21.957 ± 4.871
Standard-DPO	✓	✓	–	✓	0.678 ± 0.004	0.289 ± 0.003	9.286 ± 4.009	13.295 ± 5.072
Generic-action DPO	✓	–	✓	✓	$\underline{0.688} \pm 0.006$	$\underline{0.301} \pm 0.006$	$\underline{7.026} \pm 1.892$	$\underline{9.861} \pm 2.165$
VasoDPO	✓	✓	✓	✓	$\mathbf{0.697} \pm 0.005$	$\mathbf{0.311} \pm 0.006$	$\mathbf{0.510} \pm 0.033$	$\mathbf{0.967} \pm 0.013$
B. Additional training controls								
Context-shuffled DPO	×	✓	✓	✓	0.673 ± 0.005	0.285 ± 0.004	18.953 ± 3.627	20.367 ± 5.283
Weighted SFT	✓	n/a	✓	–	0.677 ± 0.003	0.294 ± 0.005	1.010 ± 0.904	68.369 ± 0.653

Notes. Align.: state–pair alignment; MAP: MAP guidance; Omit.: corpus-normalized omission weighting; Pair.: DPO. Dashes denote absent factors; n/a indicates inapplicability to Weighted SFT. Bold/underline: best/second-best mean in Panel A.

Matched factor comparisons. Under both weighting schemes, MAP-guided construction yields higher mean TSTR AUROC/AUPRC and lower mean CGVR/NZR than generic construction. Omission weighting produces the same pattern under both construction schemes. Shuffling patient states across preference pairs while preserving responses and weights lowers mean TSTR AUROC/AUPRC from $0.697/0.311$ to $0.673/0.285$ and raises mean CGVR/NZR from $0.510\%/0.967\%$ to $18.953\%/20.367\%$. These changes support preserving the correspondence between patient states and preference pairs during training.

Weighted SFT uses the chosen responses from the same retained corpus with the same omission weights. It achieves mean CGVR/NZR of $1.010/68.369\%$, compared with $0.510/0.967\%$ for VasoDPO. On the same retained corpus with the same omission weights, preference optimization yields substantially less concentration near zero than weighted supervised continuation. Weighted SFT also illustrates the distinction between directional violations in MAP-eligible states and concentration near zero across all parsed actions.

Complementary evaluation controls. The evaluation controls in Appendix Table 4 show how the metrics respond to restricted action variation and disrupted state–action correspondence. Constant zero achieves TSTR AUROC/AUPRC of $0.676/0.296$ with an NZR of 100%. Train median

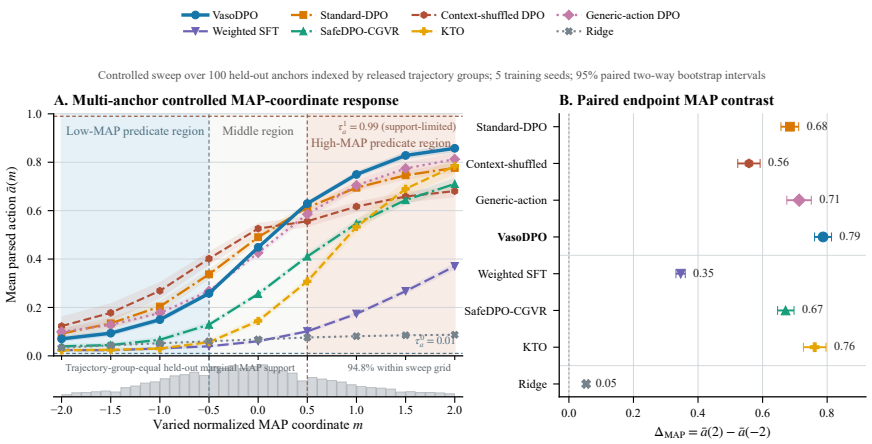

Figure 2: **Multi-anchor controlled MAP-coordinate response.** (A) Mean MAP responses for 100 model-independent anchors from distinct groups, fixed before generation, with other inputs held constant. Bands: 95% paired two-way bootstrap intervals over seeds and anchors; lines: prespecified predicates. The interval $[-2, 2]$ contains 94.8% of the group-weighted held-out MAP distribution. (B) Signed endpoint contrast $\Delta_{\mathrm{MAP}} = \bar{a}(2) - \bar{a}(-2)$, with matching intervals.

achieves zero CGVR and NZR while assigning the same action to every record. Assigning shuffled training actions to held-out states yields 476 distinct values after rounding to three decimal places, despite disrupting correspondence between actions and patient states. Constant actions can obtain favorable predictive or diagnostic scores, while shuffled actions retain diversity after disrupting their correspondence with patient states. These observations motivate examining responsiveness to patient inputs alongside the reported metrics.

Controlled multi-anchor response. Figure 2 examines how generated actions respond when MAP varies and the other inputs remain fixed. Across 100 held-out anchors and five training seeds, VasoDPO has a positive endpoint contrast $\Delta_{\mathrm{MAP}} \approx 0.79$, compared with 0.05 for Ridge. KTO also has a large endpoint difference of approximately 0.76, alongside a mean NZR of 43.043% on the held-out records. KTO's large endpoint contrast and high NZR show that sensitivity under the controlled sweep and concentration at recorded states capture different aspects of model behavior.

Weight and threshold sensitivity. In the seed-42 weight sweep, $\omega_{\mathrm{omit}} = 5$ achieves the highest AUROC/AUPRC and lowest CGVR/NZR among tested values (Appendix Table 7). With policies fixed, averaging across the evaluation-threshold grids and five training seeds yields LMW/NZR of $1.035\%/1.267\%$ for VasoDPO versus $18.484\%/13.295\%$ for Standard-DPO (Appendix Table 8).

6 Discussion and Limitations

Matched comparisons support the contributions of MAP guidance and omission weighting under the fixed protocol. The evaluation controls show that favorable predictive scores and low violation rates can coexist with concentrated action outputs, while diverse actions can lack correspondence with patient states. These findings support joint assessment of predictive utility, action distributions, and responsiveness to patient inputs.

The main comparison uses one retrospective cohort, one split with disjoint trajectory groups, five training seeds, and one construction seed. Some configuration choices were informed during the evaluation process; independent evaluation with configurations fixed in advance remains an important robustness check. The diagnostic protocol shares the MAP threshold family used in preference construction, and the upper action threshold has limited empirical support in high-MAP states. Future evaluation should cover additional cohorts, construction seeds, independently specified diagnostic criteria, and patient-level generalization, together with broader direct regression comparisons and assessment of clinical utility and causal effects.

AI Use Statement Generative AI tools were used only for editorial and presentation assistance, including language polishing, LaTeX formatting, consistency checking, and limited feedback on figure layout and clarity. The authors independently formulated the research questions and VasoDPO methodology, including the state-fixed preference construction and omission-based weighting scheme, and carried out the preprocessing, model implementation, experimental design, statistical analysis, and scientific interpretation. Generative AI was not used to generate experimental values or simulated results. Any AI-assisted edits were manually reviewed and cross-checked against the source code, fixed experimental configurations and artifacts, and cited literature. The authors take full responsibility for the final manuscript, analyses, claims, code, and figures.

Ethics. We accessed the processed cohort through PhysioNet as credentialed users under the applicable license and Data Use Agreement after completing the required training. The anonymized supplement contains source code, experiment configurations, and reproducibility utilities in accordance with the access restrictions of the source cohort. The normalized action coordinates and diagnostic thresholds are study-specific quantities defined for the retrospective evaluation protocol. Prospective validation and appropriate clinical governance are necessary for future clinical translation.

Reproducibility. Sections 3 and 4, together with Appendices A through D, specify the state representation, state-fixed preference construction, minimum action-gap filtering, corpus-normalized omission weighting, group-disjoint data splitting procedure, optimization settings, evaluators, parser, and controlled MAP-response protocol. An anonymized source-code supplement is provided with the submission, including implementations of state projection and serialization, MAP-guided and matched generic-action preference construction, the principal training and baseline pipelines used in the main experiments, action generation, action-replacement TSTR evaluation, direct action diagnostics, and the multi-anchor controlled MAP audit. It further includes experiment configuration files, preprocessing and reporting entry points, environment and preflight checks, and data-independent unit tests for key pipeline invariants. The corresponding codebase will be released publicly upon acceptance. Reproducing the full experiments requires credentialed access to the source clinical cohort and the required base-model weights under their respective access terms.

REFERENCES

Harry Amad, Zhaozhi Qian, Dennis Frauen, Julianna Piskorz, Stefan Feuerriegel, and Mihaela van der Schaar. Improving the generation and evaluation of synthetic data for downstream medical causal inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Oishi Banerjee, Hong-Yu Zhou, Kay Wu, Subathra Adithan, Stephen Kwak, and Pranav Rajpurkar. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan O. Arik. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. In *International Conference on Learning Representations*, 2025.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pp. 286–305. PMLR, 2017.

Zizhen Deng, Wei Wu, Chi Zhang, Xitong Zhao, Minghao Pu, Yanbin Bu, Yanfeng Liao, Changguan Wang, Jiarui Yang, Yanni Wang, et al. A generalist precision medication framework using temporal causal inference based on treatment-free physiological profiles. *Nature Communications*, 17:6562, 2026.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12634–12651. PMLR, 2024.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. *Critical Care Medicine*, 49(11): e1063–e1143, 2021.

Oscar Freyer, Rebecca Mathias, Hannah Sophie Muti, Henry Orlovsky, Stephan Buch, Max Ostermann, Anett Schönfelder, Akira-Sebastian Poncette, Adel Bassily-Marcus, and Stephen Gilbert. The regulation of artificial intelligence in intensive care units: From narrow tools to generalist systems. *npj Digital Medicine*, 9(1):246, 2026.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Mengxuan Hu, Vivek V. Datla, Anoop Kumar, Zihan Guan, Sheng Li, Alfy Samuel, and Daben Liu. Alignment-weighted DPO: A principled reasoning approach to improve safety alignment. In *International Conference on Learning Representations*, 2026.

Yuxin Jiang, Bo Huang, Yufei Wang, Xingshan Zeng, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, and Wei Wang. Bridging and modeling correlations in pairwise data for direct preference optimization. In *International Conference on Learning Representations*, 2025.

Dain Kim, Jiwoo Lee, Jaehoon Yun, Yong Hoe Koo, Qingyu Chen, Hyunjae Kim, and Jaewoo Kang. Benchmarking direct preference optimization for medical large vision–language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5052–5067. Association for Computational Linguistics, 2026a.

Geon-Hyeong Kim, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, Youngsoo Jang, and Moontae Lee. SafeDPO: A simple approach to direct preference optimization with enhanced safety. In *International Conference on Learning Representations*, 2026b.

Taeri Kim, Jiho Heo, Hyunjoon Kim, and Sang-Wook Kim. HI-DR: Exploiting health status-aware attention and an EHR graph+ for effective medication recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11950–11958, 2025.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17564–17579. PMLR, 2023.

Anton Danholt Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. SynthEval: A framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39(1):6, 2025.

Seonggyun Lee, Sungjun Lim, Seojin Park, Soeun Cheon, and Kyungwoo Song. Semi-supervised preference optimization with limited feedback. In *International Conference on Learning Representations*, 2026.

Jiajia Li, Zhouyu Guan, Jing Wang, Carol Y. Cheung, Yingfeng Zheng, Lee-Ling Lim, et al. Integrated image-based deep learning and language models for primary diabetes care. *Nature Medicine*, 30 (10):2886–2896, 2024.

Luchen Li, Matthieu Komorowski, and Aldo A. Faisal. Optimizing sequential medical treatments with auto-encoding heuristic search in POMDPs. *arXiv preprint arXiv:1905.07465*, 2019.

Anton Matsson, Lena Stempfle, Yaochen Rao, Zachary R. Margolin, Heather J. Litman, and Fredrik D. Johansson. How should we represent history in interpretable models of clinical policies? In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 714–734, 2025.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dipesh Niraula, Kyle C. Cuneo, Ivo D. Dinov, Brian D. Gonzalez, Jamalina B. Jamaluddin, Jionghua Judy Jin, et al. Intricacies of human–ai interaction in dynamic decision-making for precision oncology. *Nature Communications*, 16(1):1138, 2025.

Weijie Niu, Alberto Huertas Celdran, Karoline Siarsky, and Burkhard Stiller. FEST: A unified framework for evaluating synthetic tabular data. In *Proceedings of the 11th International Conference on Information Systems Security and Privacy*, volume 1, pp. 434–444. SciTePress, 2025.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018.

Yuan Pu, Dennis Shung, Alexander Tong, Nicole Zhang, and Yuki Kawamura. Clinical time series datasets for trajectory flow matching evaluation: ICU sepsis, ICU cardiac arrest, and ICU GIB cohorts. *PhysioNet*, March 2026. Version 1.0.0.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pp. 147–163. PMLR, 2017.

Thomas Savage, Stephen P. Ma, Abdessalem Boukil, Ekanath Rangan, Vishwesh Patel, Ivan Lopez, and Jonathan Chen. Fine-tuning methods for large language models in clinical medicine by supervised fine-tuning and direct preference optimization: Comparative evaluation. *Journal of Medical Internet Research*, 27:e76048, 2025.

Mihaela Stoian, Salijona Dyrmishi, Maxime Cordy, Thomas Lukasiewicz, and Eleonora Giunchiglia. How realistic is your synthetic data? constraining deep generative models for tabular data. In *International Conference on Learning Representations*, 2024.

Weicong Tan, Weiqing Wang, Xin Zhou, Wray Buntine, Gordon Bingham, et al. OntoMedRec: Logically-pretrained model-agnostic ontology encoders for medication recommendation. *World Wide Web*, 27:28, 2024.

Richa Verma, Bavish Kulur, Sanjay Chawla, and Balaraman Ravindran. PREFINE: Preference-based implicit reward and cost fine-tuning for safety alignment. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1004–1012. International Foundation for Autonomous Agents and Multiagent Systems, 2026.

Minghao Wu, Yuting Yan, Zhenyang Cai, Ke Ji, Chuangsen Fang, Ziying Sheng, Xidong Wang, Rongsheng Wang, Hejia Zhang, Shuang Li, Benyou Wang, and Hongyuan Zha. Agentifying patient dynamics within LLMs through interacting with clinical world model. *arXiv preprint arXiv:2605.14723*, 2026.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Junming Yang, Ning Xu, Biao Liu, Shiqi Qiao, and Xin Geng. Alignment through meta-weighted online sampling: Bridging the gap between data generation and preference optimization. In *International Conference on Learning Representations*, 2026a.

Ning Yang, Hai Lin, Yibo Liu, Baoliang Tian, Guoqing Liu, and Haijun Zhang. Token-importance guided direct preference optimization. In *International Conference on Learning Representations*, 2026b.

Jiancheng Ye, Donna Woods, Neil Jordan, and Justin Starren. The role of artificial intelligence for the application of integrating electronic health records and patient-generated data in clinical decision support. In *AMIA Joint Summits on Translational Science Proceedings*, pp. 459–467, 2024.

Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis L. Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107198–107224, 2024.

Yichi Zhang, Yue Ding, Jingwen Yang, Tianwei Luo, Dongbai Li, Ranjie Duan, Qiang Liu, Hang Su, Yinpeng Dong, and Jun Zhu. Towards safe reasoning in large reasoning models via corrective intervention. In *International Conference on Learning Representations*, 2026.

A FULL VASODPO CONSTRUCTION

This appendix specifies the state serialization, action rendering and parsing, and rejected-action construction used to form the preference corpus.

Table 3: Notation used in the method. Numeric actions and textual responses are distinguished throughout.

Symbol	Meaning
$N,\ i$	Number of training records and record index.
x_i, a_i, y_i, g_i	Clinical context, normalized recorded action, mortality outcome, and trajectory-group identifier.
Ψ, Φ, S_i	Clinical field selector, text serializer, and serialized state.
$a_i^+, a_i^-, r_i^+, r_i^-$	Chosen and rejected numeric actions and their rendered responses.
$\mathrm{RenderAction}$	Fixed response template with three-decimal action formatting.
$r_i, \mathrm{ParseAction}, q_i, \hat{a}_i$	Generated response, evaluation parser, acceptance flag, and parsed action.
$m(S_i), \tau_{\mathrm{MAP}}^-, \tau_{\mathrm{MAP}}^+$	Recorded standardized MAP and its low/high thresholds.
$\mathbf{1}\{\cdot\}, L_i, H_i$	Indicator function and low/high MAP membership indicators.
C_{MAP}, ξ	Rejected-action construction operator and fixed pseudorandom stream.
$\mathcal{I}_{\mathrm{SFT}}, \mathcal{I}_{\mathrm{pref}}, \delta_a$	Eligible training indices, retained pair indices, and minimum action gap.
$\mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{SFT}}, \mathcal{D}_{\mathrm{pref}}$	Training records, indexed supervised corpus, and weighted preference corpus.
$h_i^{\mathrm{omit}}, \omega_i, Z$	Omission indicator, raw pair weight, and corpus mean weight.
$\omega_{\mathrm{omit}}, \omega_{\mathrm{std}}$	Raw weights for omission pairs and remaining pairs.
$\pi_0, \pi_{\mathrm{SFT}}, \pi_\theta, \pi_{\mathrm{ref}}$	Base, SFT, trainable, and frozen reference policies.
$\Delta_{\theta,i}, \Delta_{\mathrm{ref},i}$	Chosen-versus-rejected log-probability margins for pair i.
σ, β, θ	Logistic sigmoid, DPO margin coefficient, and trainable policy parameters.
$\mathcal{K}, k$	Set of training seeds and an individual training seed.

A.1 PROJECTION AND SERIALIZER

The selector Ψ retains fields in this order: `apache`, `time_scaled`, `hr_normalized`, and `map_normalized`. The serializer Φ writes one labeled field per line, prefixes the fields with `Patient Clinical State:`, and appends the fixed task question shown below. This complete user message is S_i. Numeric values are formatted to three decimal places; supplied text values are preserved. Time uses the released scaled coordinate, and heart rate and MAP use the released standardized coordinates.

Missing or negative-sentinel APACHE values use the literal token `Score Not Computed (Missing Physiology Data)`. Missing time, heart rate, or MAP uses `Unrecorded / Measurement unavailable (cause unknown)`. These strings are fixed missingness labels. Training eligibility additionally requires observed MAP and action values, as specified in Appendix A.2.

The following example uses illustrative numeric values and reproduces the field labels and task question emitted by the serializer. Long lines wrap for display.

```
Patient Clinical State:
[Macro] APACHE Baseline Severity Score:  42.000
[Meso] Normalized ICU Admission Time:  0.250
[Meso] Current Heart Rate (Z-score, dynamic physiological
response):  0.100
[Meso] Current Mean Arterial Pressure (Z-score, represents
tissue perfusion under current support):  -0.800
Question:  Based on this multi-scale hemodynamic state, what
is the appropriate continuous Vasopressor (Norepinephrine)
infusion dosage [0.0 to 1.0]?
```

SFT, DPO, and VasoDPO generation use the same system message:

```
You are an expert clinical AI. Based on the multi-scale
patient state, predict the appropriate medical outcome or
intervention.
```

The tokenizer's chat template combines this system message with S_i as the user message. SFT adds the recorded response as an assistant message and uses `add_generation_prompt=False`. DPO and generation use `add_generation_prompt=True`; DPO appends the same `<|eot_id|>` terminator to both response strings. The fixed system message and chat formatting are implicit in $\pi_\theta(r \mid S_i)$. The implementation is in `src/vasodpo/data/builders/sepsis.py` (`build_prompt`), `src/vasodpo/data/serialization.py`, and the SFT, DPO, and generation modules of the implementation.

The APACHE field provides static retrospective context within a trajectory group; it is not recomputed or timestamp-aligned for individual action rows. The separate `apache_outcome_prob`, action target, mortality outcomes, and group identifier are excluded from S_i.

A.2 Action eligibility, rendering, and parsing

This subsection gives the corpus and parser definitions used by Algorithm 1.

Let $\mathcal{I}_{\mathrm{SFT}}$ contain exactly the training rows with an observed action and observed `map_normalized`; the training split contains 117,812 such rows. Other state fields may be missing and are encoded using the missingness labels in Appendix A.1. The SFT corpus is

$$\mathcal{D}_{\mathrm{SFT}} = \{(S_i, \mathrm{RenderAction}(a_i))\}_{i \in \mathcal{I}_{\mathrm{SFT}}}. \tag{A.1}$$

The preference candidate set is the same $\mathcal{I}_{\mathrm{SFT}}$ before action-gap filtering; no row enters either corpus when the action or the MAP value required by the predicates is missing.

Action rendering and parsing. The corpus builder uses a deterministic canonical response template,

$$\begin{aligned}\mathrm{RenderAction}(a) = {} & \texttt{[Micro] Continuous Vasopressor} \\ & \texttt{(Norepinephrine) Infusion Dosage:} \\ & \mathrm{format}(a, \texttt{.3f}).\end{aligned} \tag{A.2}$$

Here $\mathrm{format}(a, \texttt{.3f})$ produces a fixed-point decimal string with three digits after the decimal point. The displayed text segments are concatenated into a single response line. The same template is used for every SFT target and for both the chosen and rejected sides of each preference pair. Observed actions are read from the audited normalized column (whose training values lie in $[0, 1]$); constructed rejected actions are clipped to $[0, 1]$ before this three-decimal formatting, so sampled actions are rounded only when rendered as text. During DPO preprocessing, the prompt is chat-templated and the same `<|eot_id|>` terminator is appended to both response strings.

At evaluation time, $\mathrm{ParseAction}$ is the implementation's case-insensitive regular-expression parser. It extracts the first numeric token following either "`Dosage:`" or "`Dosage Target:`", accepting signed decimal and scientific notation, and clips the parsed value to $[0, 1]$. The second label is retained for compatibility with evaluated output schemas; the canonical training renderer above uses "`Dosage:`". If no labeled token is found, the parser accepts only a response that can be converted as a whole to a floating-point number; otherwise it returns an invalid parse. In the evaluator, $q_j = 1$ exactly when this result is not NaN; empty or malformed responses therefore have $q_j = 0$ and no $\hat{a}_j$ is used. The parser permits surrounding text when a labeled dosage token is present, uses the first labeled match when several are present, and clips finite values outside $[0, 1]$ rather than rejecting them. The literal `NaN` remains invalid; a bare infinity string accepted by the floating-point fallback is clipped to the corresponding endpoint. Parser acceptance, $|\mathcal{V}|/M$, measures the fraction of responses yielding a numeric action under these rules and is 100% in the reported artifacts. Separate counts of exact-template matches, fallback uses, and clipping events are unavailable in the current summaries.

A.3 Action-space perturbation operator

This subsection expands Equation 3 with the exact branch probabilities and intermediate-MAP fallback used to construct rejected actions. The operator perturbs the action while preserving the recorded patient context.

Seeded construction stream. Corpus construction processes the audited training rows in their fixed file order using one Python pseudorandom stream ξ initialized with seed 42. Let $\mathrm{Next}(\xi)$ return the next draw from $\mathrm{Uniform}(0,1)$. For real lower and upper bounds $l \le h$, define

$$\mathrm{Unif}_\xi(l,h) = l + (h-l)\,\mathrm{Next}(\xi). \tag{A.3}$$

For a low- or high-MAP row, $u_i = \mathrm{Next}(\xi)$ selects the point-mass or interval branch. If the interval branch is selected, Unif_ξ consumes the next draw. An intermediate-MAP interval branch consumes one draw directly; the intermediate-MAP zero branch consumes none. Under $\tau^-_{\mathrm{MAP}} = -0.5$ and $\tau^+_{\mathrm{MAP}} = 0.5$, the implementation constructs

$$C_{\mathrm{MAP}}(S_i, a_i^+; \xi) = \begin{cases} 0, & m(S_i) < \tau^-_{\mathrm{MAP}} \ \wedge \ u_i > 0.3, \\ \mathrm{Unif}_\xi(0.001, 0.020), & m(S_i) < \tau^-_{\mathrm{MAP}} \ \wedge \ u_i \le 0.3, \\ 1, & m(S_i) > \tau^+_{\mathrm{MAP}} \ \wedge \ u_i > 0.3, \\ \mathrm{Unif}_\xi(0.70, 0.99), & m(S_i) > \tau^+_{\mathrm{MAP}} \ \wedge \ u_i \le 0.3, \\ \mathrm{Unif}_\xi(0.20, 0.60), & \tau^-_{\mathrm{MAP}} \le m(S_i) \le \tau^+_{\mathrm{MAP}} \ \wedge \ a_i^+ \le 0.05, \\ 0, & \tau^-_{\mathrm{MAP}} \le m(S_i) \le \tau^+_{\mathrm{MAP}} \ \wedge \ a_i^+ > 0.05. \end{cases} \tag{A.4}$$

All constructed actions are clipped to $[0,1]$, and the preference corpus retains a pair only when $|a_i^+ - a_i^-| \ge \delta_a$ with $\delta_a = 0.1$. The MAP predicates select the construction branch, and the action-gap filter selects the retained comparisons. In the intermediate MAP group, records with $0.05 < a_i^+ < 0.1$ produce a zero alternative and are removed by this filter. The interval-branch probability 0.3 (point-mass probability 0.7), the intermediate action cutoff 0.05, the action intervals, and $\delta_a = 0.1$ are fixed constants in `src/vasodpo/data/builders/sepsis.py`. The MAP thresholds use standardized MAP coordinates; the action cutoffs and intervals use normalized action coordinates. The same constants construct the canonical corpus reused across the reported training seeds.

The canonical preference corpus is constructed once with this seeded stream and then frozen. VasoDPO and methods defined on the canonical pairs use the same serialized `dpo.jsonl` for optimization seeds $\{42, 43, 44, 45, 46\}$; changing the optimization seed does not regenerate rejected actions. Context-shuffled and generic-action DPO use their stated transformed corpora, each of which is also fixed across its reported optimization seeds. The corpus construction seed is therefore distinct from a model training seed. The canonical manifest records construction seed 42, 117,812 input rows, and 76,442 retained pairs.

A.4 Pair eligibility and preference assignment

The action-gap filter selects the preference pairs, and the recorded actions define their chosen responses.

The preference index set applies the minimum action gap used in the main text:

$$\mathcal{I}_{\mathrm{pref}} = \{i \in \mathcal{I}_{\mathrm{SFT}} : |a_i^+ - a_i^-| \ge \delta_a\}. \tag{A.5}$$

The weighted preference corpus is

$$\mathcal{D}_{\mathrm{pref}} = \{(S_i, r_i^+, r_i^-, \omega_i)\}_{i \in \mathcal{I}_{\mathrm{pref}}}. \tag{A.6}$$

For each retained pair, the recorded action $a_i^+ = a_i$ defines the chosen response, and the constructed alternative defines the rejected response. This assignment specifies the training preference $r_i^+ \succ_{S_i} r_i^-$ under the shared serialized state S_i.

The implementation-level omission indicator and sample weight are

$$\begin{aligned} h_i^{\mathrm{omit}} &= \mathbf{1}\{a_i^+ > 0 \wedge a_i^- = 0\}, \\ \omega_i &= \omega_{\mathrm{omit}} h_i^{\mathrm{omit}} + \omega_{\mathrm{std}}(1 - h_i^{\mathrm{omit}}). \end{aligned} \tag{A.7}$$

For retained pairs with $a_i^- = 0$, the action-gap criterion implies $a_i^+ \ge \delta_a$. These omission labels determine the pair weights used in Appendix B.

B Optimization Details

Pair weights determine the relative contributions of preference comparisons. Corpus normalization sets their mean weight to one in Eq. 8.

B.1 Omission weighting and corpus normalization

For a fixed preference corpus, $Z = |\mathcal{I}_{\mathrm{pref}}|^{-1} \sum_{i \in \mathcal{I}_{\mathrm{pref}}} \omega_i$ is constant with respect to θ. The normalized weights have mean one, so changing ω_{omit} changes the relative contributions of retained pairs within the logistic DPO objective. We compute Z once from the retained corpus and reuse the normalized weights in every optimization batch.

B.2 Gradient and effective-pair-weight interpretation

Writing $d_i = \Delta_{\theta,i} - \Delta_{\mathrm{ref},i}$, the single-pair loss $\ell_i = -(\omega_i/Z) \log \sigma(\beta d_i)$ has

$$\frac{\partial \ell_i}{\partial \Delta_{\theta,i}} = -\frac{\beta \omega_i}{Z} \sigma(-\beta d_i). \tag{B.1}$$

At a fixed preference margin, a positive pair weight rescales the loss derivative. The ratio of two normalized weights is ω_i/ω_j, so the common normalizer Z preserves their relative emphasis. Equivalently, define

$$\widetilde{w}_i = \frac{\omega_i}{\sum_{j \in \mathcal{I}_{\mathrm{pref}}} \omega_j}, \qquad \sum_{i \in \mathcal{I}_{\mathrm{pref}}} \widetilde{w}_i = 1. \tag{B.2}$$

The corpus objective is

$$\mathcal{L}_{\mathrm{VasoDPO}}(\theta) = -\sum_{i \in \mathcal{I}_{\mathrm{pref}}} \widetilde{w}_i \log \sigma(\beta d_i). \tag{B.3}$$

Thus, ω_i/Z has mean one, while $\widetilde{w}_i$ sums to one and includes the outer corpus average in Eq. 8.

B.3 SFT objective

Let $\mathcal{B} \subseteq \mathcal{I}_{\mathrm{SFT}}$ index an SFT minibatch, and let $\mathcal{T}_i$ contain the response-token positions whose next-token labels remain unmasked after completion-only collation. Write $r^+_{i,t}$ for the response token at position t and $r^+_{i,<t}$ for its preceding response tokens. Prompt and padding targets are excluded. Define the mean response loss

$$\ell_i^{\mathrm{SFT}}(\theta) = -\frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \log \pi_\theta(r^+_{i,t} \mid S_i, r^+_{i,<t}). \tag{B.4}$$

The SFT implementation averages over all unmasked response targets in the minibatch:

$$\mathcal{L}_{\mathrm{SFT}}(\theta; \mathcal{B}) = \frac{\sum_{i \in \mathcal{B}} |\mathcal{T}_i| \ell_i^{\mathrm{SFT}}(\theta)}{\sum_{i \in \mathcal{B}} |\mathcal{T}_i|}. \tag{B.5}$$

The collation mask determines which special-token targets contribute to $\mathcal{T}_i$. The resulting SFT policy initializes preference optimization. DPO uses the response log-probability sums in Eq. 7.

B.4 VasoDPO optimization pseudocode

In Algorithm 1, $\mathrm{PRNG}(42)$ initializes the construction stream with seed 42, and e indexes training epochs. Each $\mathcal{B}$ denotes an effective optimization batch, assembled across distributed processes and gradient-accumulation steps using the settings in Appendix D.5.

Algorithm 1 VasoDPO Training Procedure

Require: Training split $\mathcal{D}_{\text{train}} = \{(x_i, a_i, y_i, g_i)\}_{i=1}^{N}$
Require: Field selector Ψ, serializer Φ, operator C_{MAP}, and action gap δ_a
Require: Base policy π_0, training seed k, DPO coefficient β, and positive weights $\omega_{\text{omit}} > \omega_{\text{std}} > 0$
Require: Construction stream $\xi \leftarrow \text{PRNG}(42)$
Ensure: Trained policy π_θ

Initialize indexed corpora $\mathcal{D}_{\text{SFT}}$, $\mathcal{D}_{\text{pref}}$ and index set $\mathcal{I}_{\text{pref}} \leftarrow \emptyset$
for all record indices $i = 1, \ldots, N$ in fixed file order **do**
  if a_i or the recorded MAP value is missing **then**
    continue
  end if
  $S_i \leftarrow \Phi(\Psi(x_i))$ ▷ exclude action, outcome, and trajectory-group identifier
  $a_i^+ \leftarrow a_i$
  $r_i^+ \leftarrow \text{RenderAction}(a_i^+)$
  Append (S_i, r_i^+) to $\mathcal{D}_{\text{SFT}}$ with index i
  $a_i^- \leftarrow C_{\text{MAP}}(S_i, a_i^+; \xi)$ ▷ change the action while keeping S_i fixed
  if $|a_i^+ - a_i^-| < \delta_a$ **then**
    continue
  end if
  $r_i^- \leftarrow \text{RenderAction}(a_i^-)$
  $h_i^{\text{omit}} \leftarrow \mathbf{1}\{a_i^+ > 0 \wedge a_i^- = 0\}$
  $\omega_i \leftarrow \omega_{\text{omit}} h_i^{\text{omit}} + \omega_{\text{std}}(1 - h_i^{\text{omit}})$
  Append $(S_i, r_i^+, r_i^-, \omega_i)$ to $\mathcal{D}_{\text{pref}}$ with index i
  $\mathcal{I}_{\text{pref}} \leftarrow \mathcal{I}_{\text{pref}} \cup \{i\}$
end for
Set training seed k
$\pi_{\text{SFT}} \leftarrow \text{SFT}(\pi_0, \mathcal{D}_{\text{SFT}})$
Initialize $\pi_\theta \leftarrow \pi_{\text{SFT}}$ and freeze $\pi_{\text{ref}} \leftarrow \pi_{\text{SFT}}$
$Z \leftarrow |\mathcal{I}_{\text{pref}}|^{-1} \sum_{i \in \mathcal{I}_{\text{pref}}} \omega_i$
for $e = 1, 2, 3$ **do**
  for all effective batches of indices $\mathcal{B} \subseteq \mathcal{I}_{\text{pref}}$ **do**
    for all $i \in \mathcal{B}$ **do**
      $\Delta_{\theta,i} \leftarrow \log \pi_\theta(r_i^+ \mid S_i) - \log \pi_\theta(r_i^- \mid S_i)$
      $\Delta_{\text{ref},i} \leftarrow \log \pi_{\text{ref}}(r_i^+ \mid S_i) - \log \pi_{\text{ref}}(r_i^- \mid S_i)$
      $\ell_i \leftarrow -(\omega_i / Z) \log \sigma(\beta[\Delta_{\theta,i} - \Delta_{\text{ref},i}])$
    end for
    Take an optimizer step on $|\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \ell_i$
  end for
end for
return π_θ

C Complete Experimental Protocol

This appendix specifies the data partition, downstream evaluator, action diagnostics, and controlled MAP audit used in Section 4.

C.1 Dataset and HADM_ID-group-disjoint split

The processed PhysioNet release contains 147,212 action-associated records from 3,362 HADM_ID trajectory groups, all within the first 24 hours after ICU admission. Its `label` field assigns 2,689 groups to the released training partition, 336 to validation, and 337 to testing. We pool the three released partitions and construct the study split from all 3,362 groups. Using seed 42, we stratify groups by in-hospital mortality and assign each HADM_ID group to one partition while targeting an 80/20 ratio of training to test records. The resulting training partition contains 117,812 records from 2,698 groups, and the test partition contains 29,400 records from 664 groups, with no HADM_ID overlap. We update the `label` field to record these study-specific assignments. The training partition supplies the SFT and preference corpora and the states used to generate actions for TSTR training. The test partition supplies the real records for TSTR evaluation and the states used for action diagnostics.

The splitter in `src/vasodpo/data/split.py` assigns whole groups within each mortality stratum. It orders groups by a SHA-256 digest of the seed, stratum, and group key, and selects the nonempty proper prefix whose row count is closest to 20% of the stratum's records. A tie selects the shorter prefix; a stratum containing one group remains in training. Group outcomes must be consistent across their nonmissing labels. The saved split manifest records the resulting counts and verifies zero group overlap.

We retain the released `time_scaled`, `hr_normalized`, `map_normalized`, and `norepi_inf_scaled` values. These transformations precede the study split, and their parameters are inherited from the processed release. All compared methods therefore use the same input coordinates and action scale. The release provides `HADM_ID` as its trajectory identifier but does not include the native eICU person, hospital-stay, or unit-stay identifiers. The APACHE predicted-mortality field (`apache_outcome_prob`) is used only to construct the internal severity cohort shift and is excluded from the serialized model context and the TSTR predictor.

C.2 TSTR DATASETS AND EVALUATOR

The `synthetic_train` artifact retains training states and mortality outcomes and replaces the norepinephrine column with generated actions. The downstream predictor uses `apache`, `time_scaled`, `hr_normalized`, `map_normalized`, and the action column to predict `HOSP_MORT`. Evaluation uses the real test records and their recorded actions. The separate `generated_test` artifact contains actions generated for held-out states and supplies the direct action diagnostics.

The shared XGBoost classifier uses 300 trees, maximum depth 6, learning rate 0.05, row and column subsampling rates of 0.8, the histogram tree method, log-loss evaluation, and random seed 42. Preprocessing is fitted on the training partition: numerical features receive median imputation, and categorical features, if present, receive most-frequent imputation and one-hot encoding with unseen categories ignored. Records require an observed mortality label and a valid action. The state-only comparison omits the action feature. Across action-model training seeds, the evaluator configuration and test records remain fixed.

C.3 COMPLETE DIRECTIONAL DIAGNOSTIC DEFINITIONS

This subsection provides the full parse-failure handling and normalization behind CGVR, LMW, HME, and NZR in the main tables.

For an evaluation artifact with M records indexed by $j = 1, \dots, M$, let r_j be the response generated for state S_j. As in Equation 1, the parser returns acceptance flag q_j and, when $q_j = 1$, action $\hat{a}_j$:

$$(q_j, \hat{a}_j) = \mathrm{ParseAction}(r_j), \qquad \mathcal{V} = \{j \in \{1, \dots, M\} : q_j = 1\}. \tag{C.1}$$

Parser acceptance is $|\mathcal{V}|/M$. NZR and numeric action summaries use records in $\mathcal{V}$. Directional rates use all records in their respective MAP eligibility groups and count parse failures as violations. When $q_j = 0$, no numeric action comparison is evaluated.

All 29,400 held-out records have finite MAP values: 8,949 are in the low-MAP group, 12,989 in the intermediate group, and 7,462 in the high-MAP group under the primary thresholds. Successfully parsed actions from the intermediate group enter NZR.

Using the indicators in Eq. 2, define $\mathcal{E}_{\mathrm{low}} = \{j : L_j = 1\}$, $\mathcal{E}_{\mathrm{high}} = \{j : H_j = 1\}$, and $\mathcal{E}_{\mathrm{crit}} = \mathcal{E}_{\mathrm{low}} \cup \mathcal{E}_{\mathrm{high}}$. The strict indicators are the same d_j^{low} and d_j^{high} used in the main text:

$$\begin{aligned} d_j^{\mathrm{low}} &= \mathbf{1}\{q_j = 0 \vee (q_j = 1 \wedge \hat{a}_j \le \tau_a^0)\}, \quad j \in \mathcal{E}_{\mathrm{low}}, \\ d_j^{\mathrm{high}} &= \mathbf{1}\{q_j = 0 \vee (q_j = 1 \wedge \hat{a}_j \ge \tau_a^1)\}, \quad j \in \mathcal{E}_{\mathrm{high}}. \end{aligned} \tag{C.2}$$

The action thresholds are $\tau_a^0 = 0.01$ and $\tau_a^1 = 0.99$. LMW is

$$\mathrm{LMW} = \frac{\sum_{j \in \mathcal{E}_{\mathrm{low}}} d_j^{\mathrm{low}}}{|\mathcal{E}_{\mathrm{low}}|}. \tag{C.3}$$

The remaining rates are

$$\begin{aligned} \mathrm{CGVR} &= \frac{\sum_{j\in\mathcal{E}_{\mathrm{low}}} d_j^{\mathrm{low}} + \sum_{j\in\mathcal{E}_{\mathrm{high}}} d_j^{\mathrm{high}}}{|\mathcal{E}_{\mathrm{crit}}|}, \\ \mathrm{HME} &= \frac{\sum_{j\in\mathcal{E}_{\mathrm{high}}} d_j^{\mathrm{high}}}{|\mathcal{E}_{\mathrm{high}}|}, \\ \mathrm{NZR} &= \frac{1}{|\mathcal{V}|}\sum_{j\in\mathcal{V}} \mathbf{1}\{\hat{a}_j \le \tau_a^0\}. \end{aligned} \tag{C.4}$$

These proportions are reported as percentages. An empty denominator set produces a missing rate, stored as JSON `null`. Parse failures contribute to directional rates and are excluded from numeric action summaries. The five-seed aggregation reports a metric only when it is available for all five seeds.

C.4 Controlled MAP response audit

Anchor selection. The anchor manifest is constructed independently of model outputs and diagnostic scores. Eligible held-out contexts are stratified by quintiles of the APACHE-derived severity covariate and scaled time, yielding $5 \times 5 = 25$ non-MAP context strata. APACHE-severity quintiles are defined across released trajectory groups. Scaled-time quintiles are computed over held-out rows with weight $1/(Gn_g)$ for a row in released trajectory group g, where $G = 664$ is the number of held-out trajectory groups and n_g is the number of held-out rows in group g. Thus, each released trajectory group contributes equal total mass to the time stratification. Using anchor-selection seed 2027, we sample four eligible contexts per stratum subject to a global constraint allowing one anchor from each released trajectory group, yielding 100 anchors from 100 distinct released trajectory groups. If multiple eligible rows from a selected group fall within the relevant stratum, one row is chosen using a deterministic seeded rule. Anchor selection uses neither observed actions, outcomes, generated actions, model-specific diagnostic scores, nor controlled-response results. The resulting manifest is frozen before controlled generation and reused unchanged across all methods, training seeds, and MAP-grid values.

Controlled coordinate sweep. Let $\mathcal{H}$ denote the resulting 100-anchor manifest. Each anchor is recorded with its released trajectory-group identifier and standardized MAP coordinate $m :=$ `map_normalized`. We use the fixed grid

$$\mathcal{G}_m = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}.$$

Within an anchor $h \in \mathcal{H}$, the APACHE-derived severity, scaled time, standardized heart rate, missingness representation, prompt template, and decoding configuration remain fixed; only m changes across the nine serialized prompts. Let $\mathcal{K} = \{42, 43, 44, 45, 46\}$ denote the training seeds, and let $\hat{a}_{h,k}(m)$ be the parsed action generated for anchor h by the model trained with seed k, with MAP set to m. We summarize the responses by the mean action curve and its signed endpoint difference:

$$\begin{aligned} \bar{a}(m) &= \frac{1}{|\mathcal{H}||\mathcal{K}|}\sum_{h\in\mathcal{H}}\sum_{k\in\mathcal{K}} \hat{a}_{h,k}(m), \\ \Delta_{\mathrm{MAP}} &= \frac{1}{|\mathcal{H}||\mathcal{K}|}\sum_{h\in\mathcal{H}}\sum_{k\in\mathcal{K}} \left[\hat{a}_{h,k}(2) - \hat{a}_{h,k}(-2)\right]. \end{aligned} \tag{C.5}$$

Here Δ_{MAP} is signed: its magnitude summarizes endpoint sensitivity, and its sign records the response direction. These quantities describe model sensitivity to a controlled change in its MAP input. Uncertainty uses 2,000 paired two-way bootstrap replicates over seeds and anchors. Each replicate resamples training-seed and anchor indices with replacement and applies the same sampled indices to every method; the plotted intervals are the 2.5th and 97.5th percentiles. Pairwise differences use the same replicate-level contrasts. The audit artifact contains the anchor manifest and parsed outputs indexed by seed, anchor, method, and MAP value. The plotted curves use the generated means on the specified grid.

The support strip uses all 29,400 held-out records from 664 trajectory groups. Each record in group g receives weight $1/(Gn_g)$, giving each group equal total weight. The interval $[-2, 2]$ contains 94.8%

of this weighted marginal MAP distribution. This coverage describes MAP alone; joint support of the modified multivariable states requires a separate assessment.

C.5 Threshold and support sensitivity

The primary thresholds for standardized MAP and normalized actions are $(\tau^{-}_{\mathrm{MAP}}, \tau^{+}_{\mathrm{MAP}}, \tau^{0}_{a}, \tau^{1}_{a}) = (-0.5, 0.5, 0.01, 0.99)$. Panel A of Table 8 evaluates the low-MAP threshold grid $\{-0.25, -0.50, -0.75, -1.00\}$ crossed with the near-zero threshold grid $\{0.005, 0.010, 0.020\}$. LMW depends on both threshold coordinates, whereas NZR depends only on the near-zero threshold. Accordingly, the arithmetic mean for LMW uses 12 threshold combinations across five seeds (60 evaluations), and the arithmetic mean for NZR uses the three distinct near-zero thresholds across the same five seeds (15 evaluations). Panel B fixes $\tau^{+}_{\mathrm{MAP}} = 0.5$ and audits high-action thresholds $\{0.300, 0.500, 0.700, 0.990\}$ within high-MAP-eligible states. The trained policies and preference corpora remain fixed throughout this evaluator-threshold analysis.

C.6 Seed aggregation and uncertainty

The main aggregates use seeds $\{42, 43, 44, 45, 46\}$ on the single fixed split. A seed is included only when its checkpoint, `synthetic_train`, `generated_test`, strict protocol audit, and JSON report are present. Means and sample standard deviations summarize optimization-seed variation on the fixed split. In the seed-uncertainty figure, the 95% interval is $\bar{v} \pm t_{0.975,4} s_v/\sqrt{5}$, where $\bar{v}$ and s_v are the mean and sample standard deviation of a metric across the five seeds and $t_{0.975,4}$ is the Student t quantile with four degrees of freedom.

D Training and Baseline Configurations

This appendix specifies initialization, objectives, and execution settings for the methods in Tables 1 and 2.

D.1 SFT and VasoDPO core settings

All fine-tuned LLM policies ultimately derive from `Meta-Llama-3-8B-Instruct` and use LoRA with rank 16, scaling parameter 32, and dropout 0.05. The SFT and DPO adapters target the attention projections `q_proj`, `k_proj`, `v_proj`, and `o_proj`, and the feed-forward projections `gate_proj`, `up_proj`, and `down_proj`, with adapter bias disabled. For each seed, first-stage SFT uses learning rate 2×10^{-5} to produce $\pi^{(k)}_{\mathrm{SFT}}$. Weighted SFT and all preference-optimization runs then use the corresponding seed-matched SFT checkpoint as their initialization. VasoDPO uses $\beta = 0.1$, $(\omega_{\mathrm{omit}}, \omega_{\mathrm{std}}) = (5, 1)$, learning rate 2×10^{-6}, three epochs, and greedy decoding. The primary seed set is $\{42, 43, 44, 45, 46\}$. Appendix D.5 gives batch sizes, sequence limits, optimizer settings, numerical precision, and checkpoint selection.

D.2 Weighted supervised continuation

For each seed k, Weighted SFT starts from the same $\pi^{(k)}_{\mathrm{SFT}}$ used to initialize VasoDPO. It performs continued supervised fine-tuning on the chosen-response projection of the 76,442 records retained in $\mathcal{I}_{\mathrm{pref}}$. Each chosen response inherits the omission indicator and weight of its corresponding preference pair. Its objective is

$$\mathcal{L}_{\mathrm{WSFT}}(\theta) = \frac{1}{|\mathcal{I}_{\mathrm{pref}}|} \sum_{i \in \mathcal{I}_{\mathrm{pref}}} \frac{\omega_i}{Z} \ell^{\mathrm{SFT}}_i(\theta). \tag{D.1}$$

Here ℓ^{SFT}_i is the mean response-token loss defined in Appendix B.3. Weighted SFT uses the same retained records, raw weights, and corpus normalizer Z as VasoDPO. The implementation averages token losses within each response before applying its normalized sample weight. Weighted SFT shares VasoDPO's seed-specific initialization, retained records, chosen responses, omission labels, and $(5, 1)$ weights. It uses the supervised learning rate and batch configuration in Appendix D.5; its objective evaluates the chosen responses only.

D.3 PREFERENCE-OPTIMIZATION BASELINES

Standard-DPO. Standard-DPO uses the same 76,442 state-fixed pairs and seed-matched SFT initialization as VasoDPO, uniform pair weights $(1, 1)$, learning rate 2×10^{-6}, $\beta = 0.1$, three epochs, and the VasoDPO batch and accumulation settings.

SimPO. SimPO starts from the seed-matched SFT checkpoint and uses the same 76,442 underlying preference pairs. It uses learning rate 5×10^{-7}, scale $\beta = 2.0$ and target reward margin $\gamma = 1.0$ ($\gamma/\beta = 0.5$), and three epochs.

KTO. KTO starts from the seed-matched SFT checkpoint and converts the same 76,442 underlying pairs into 152,884 method-specific desirable/undesirable examples. It uses learning rate 5×10^{-6}, $\beta = 0.05$, desirable and undesirable weights both equal to 1, and three epochs; the audited loader uses `dataloader_drop_last=True`.

SafeDPO-CGVR. This baseline applies the primary CGVR predicates to preference training. The SafeDPO-CGVR baseline uses learning rate 2×10^{-6}, $\beta = 0.1$, and margin coefficient $\delta_{\text{safe}} = 5$. The implementation in `src/vasodpo/training/safedpo.py` marks a response as violating the rule when its parsed action is near zero in a low-MAP state or near the upper bound in a high-MAP state, using the primary thresholds. Pairs with two violating responses are discarded. If only the original chosen response violates the rule, the two responses are exchanged.

For this baseline, let v_i^+ and v_i^- be the binary violation indicators after this transformation, and let $\Delta_{\theta,i}^{\text{safe}}$ and $\Delta_{\text{ref},i}^{\text{safe}}$ be the log-probability differences in Eq. 7 computed on the transformed responses. The per-pair loss is

$$\ell_i^{\text{safe}} = -\log \sigma\big(\beta[\Delta_{\theta,i}^{\text{safe}} - \Delta_{\text{ref},i}^{\text{safe}}] - \delta_{\text{safe}}(v_i^- - v_i^+)\big)\,. \tag{D.2}$$

Training averages this loss over retained pairs.

Matched DPO configurations. The matched generic construction preserves each retained pair's state, chosen action, retention decision, absolute action gap, and omission class, while selecting a feasible rejected value without consulting MAP. When no distinct matched alternative is feasible, the canonical rejected value is retained and the fallback is recorded in the construction manifest. The MAP-guided construction uses the operator in Eq. A.4.

Generic-action DPO-U and Generic-action DPO share the generic preference corpus and use uniform and omission weights, respectively. Standard-DPO and VasoDPO share the MAP-guided preference corpus and use the same two weighting schemes. Uniform weighting assigns unit weight to every pair; omission weighting follows Eq. 5. These comparisons evaluate weighting within each fixed corpus and construction across the two corpora.

Context-shuffled DPO evaluates state–pair alignment by reassigning patient states across VasoDPO's preference pairs while keeping each chosen response, rejected response, and pair weight together. The four core variants and this control use the corresponding seed-matched SFT initialization and share the optimizer, learning rate, β, epoch count, batch settings, and decoding configuration specified in Appendix D.5.

D.4 INFERENCE-TIME AND TABULAR CONTROLS

Rule-rerank SFT samples eight responses with temperature 0.7 and top-p 0.9. The candidate penalty adds 100 for low-MAP withholding, 100 for high-MAP over-escalation, and 50 when standardized heart rate exceeds 1.5 and the action exceeds 0.8. A parse failure receives penalty 10^6. The lowest-penalty candidate is selected, with ties resolved by generation order; if all candidates fail parsing, the first is retained as an invalid response. These rules use state fields and generated actions. Its inference budget is eight sampled candidates per state.

Ridge uses median imputation, StandardScaler, and a Ridge regularization coefficient of 1.0. The XGBoost control uses 500 estimators, maximum depth 6, learning rate 0.05, subsample 0.8, and `colsample_bytree=0.8`, together with `objective=reg:squarederror`, `tree_method=hist`, and `random_state=seed`. This is distinct from the fixed TSTR mortality

evaluator, which uses 300 estimators, depth 6, and learning rate 0.05. TabICLv2 uses the tabular regressor implementation in `src/vasodpo/baselines/action_regression.py`, which loads the `tabicl-regressor-v2-20260212.ckpt` checkpoint locally and sets its random state to the training seed.

D.5 Training procedure and execution configuration

The SFT corpus includes all eligible observed actions before action-gap filtering. The preference corpus retains only pairs satisfying the minimum action gap and attaches the omission weight used by the DPO stage. Weighted SFT uses the chosen-response projection of this retained corpus and inherits the same per-record omission labels and weights. On this cohort, 76,442 of 117,812 SFT-eligible candidate rows are retained after the minimum action-gap filter; 17,381 (22.7%) are strict positive-to-zero omission contrasts weighted by ω_{omit}, with the remainder weighted by ω_{std}.

First-stage SFT and the Weighted SFT continuation use per-device batch size 4, gradient accumulation 4, six training GPUs/processes, effective batch size 96, three epochs, learning rate 2×10^{-5}, and maximum sequence length 1536. Generic-action DPO-U, Generic-action DPO, Standard-DPO, VasoDPO, and the additional Context-shuffled DPO control use per-device batch size 4, gradient accumulation 8, six training GPUs/processes, effective batch size 192, three epochs, learning rate 2×10^{-6}, maximum prompt length 1024, and maximum total sequence length 2048.

Scope of hyperparameter matching. The four core matched DPO configurations in Appendix D.3 share initialization, corpus size, optimizer, batch and accumulation settings, learning rate, β, epoch count, and decoding. Within each construction, the uniform and omission variants use identical response pairs. Cross-objective baselines use objective-specific learning rates and execution settings, as specified in their configuration descriptions.

All fine-tuned LLM runs use `adamw_torch`, a cosine learning-rate scheduler, bf16 computation, and gradient checkpointing. The reported checkpoint is the final merged state after three training epochs.

D.6 Software and hardware environment

Each training run uses six GPUs on a server with eight NVIDIA L20 GPUs, each exposing 46,068 MiB of device memory. The recorded driver version is 550.163.01, with CUDA driver compatibility 12.4. The available environment records list Python 3.10, PyTorch 2.3, and Triton 2.3.0.

E Supporting Results

Unless otherwise specified, these analyses use the main data split and evaluation protocol. The internal cohort-shift experiments use the partitions defined in Appendix E.2.

E.1 Diagnostic shortcut controls

We construct four controls directly from the training action pool and evaluate them with seed 42 under the same TSTR and action-diagnostic protocols. For each control, we form both the action-replacement training records and the generated actions for held-out states. The pool contains finite training values of `norepi_inf_scaled`, clipped to $[0, 1]$ in their original record order.

Constant and median controls. Constant zero assigns action zero to every record. Train median assigns the median of all actions in the training pool, including zeros, to every record. The Low-MAP median rule assigns the median of strictly positive training actions, floored at 0.02, when standardized MAP is below -0.5; it assigns 0.02 otherwise. If the positive pool is empty, the rule uses 0.02 throughout. In the evaluated artifacts, Train median emits 0.045863, and the Low-MAP median rule emits 0.078547 in the low-MAP group and 0.02 elsewhere.

Shuffled-action control. The Shuffled actions control permutes the training action pool and assigns the resulting values to target records in their file order, independently of the target states. The

construction uses separate Python pseudorandom streams for the training and test artifacts, initialized with seeds 42,022,234 and 42,022,344, respectively. If more outputs are needed than the pool contains, independently shuffled copies are appended; the resulting sequence is truncated to the required length. An unchanged shuffle of a pool with more than one value is rotated by one position. Both artifacts draw exclusively from the training action pool. This control changes the assignment of actions during evaluation; Context-shuffled DPO instead changes the states assigned to response pairs during preference training.

All control actions are rendered to six decimal places and passed through the shared evaluator. Unique@3dp counts distinct parsed values after rounding to three decimal places. These fixed seed-42 constructions produce the entries in Table 4.

Table 4: **Diagnostic shortcut controls (seed 42).** TSTR AUROC/AUPRC, CGVR, and NZR use the same fixed evaluator and normalized stress-test protocol as the main tables. Unique@3dp counts distinct parsed actions after rounding to three decimals; the Low-MAP median rule yields two distinct rounded action values. Bold/underline mark the best/second-best displayed scores per metric among controls; ties share a rank.

Control	AUROC ↑	AUPRC ↑	CGVR (%) ↓	NZR (%) ↓	Unique@3dp
Constant zero	<u>0.676</u>	**0.296**	54.530	100.000	1
Train median	<u>0.676</u>	**0.296**	**0.000**	**0.000**	1
Shuffled actions	<u>0.676</u>	<u>0.295</u>	<u>21.406</u>	<u>39.211</u>	476
Low-MAP median	**0.677**	<u>0.295</u>	**0.000**	**0.000**	2

E.2 Internal cohort-shift diagnostics

We evaluate two internal cohort shifts within the processed eICU cohort using training seed 42.

For the MAP-burden shift, each `HADM_ID` group is scored by the fraction of its rows with standardized MAP below -0.75; groups at or below the 70th percentile (score ≤ 0.2632; 2,358 groups and 106,964 rows) form the training partition, while those at or above the 80th percentile (score ≥ 0.3333; 741 groups and 26,667 rows) form the shifted evaluation partition. For the severity shift, the score is the trajectory-group mean APACHE-predicted mortality risk; the corresponding 70th-percentile training partition has 2,359 groups and 100,785 rows (score ≤ 0.2305), and the 80th-percentile evaluation partition has 701 groups and 32,993 rows (score ≥ 0.2808). Groups between the two cutoffs are buffered, and the training and evaluation `HADM_ID` groups do not overlap. All methods use the corresponding shift-specific training partition and are evaluated on its disjoint high-tail partition; learned checkpoints are retrained on each shift-specific training partition.

VasoDPO has the highest AUROC and AUPRC in both shifts. In the severity shift, TabICLv2 and SafeDPO-CGVR attain lower CGVR and NZR, and KTO attains lower CGVR.

E.3 Five-seed generation audit

Table 6 expands the main comparison with action means, action standard deviations, and parser acceptance.

E.4 Seed uncertainty

Figure 3 shows the uncertainty of the matched training comparisons.

E.5 Omission-weight sweep

The main experiments use $\omega_{\text{omit}} = 5$. Table 7 reports an omission-weight sensitivity analysis with training seed 42 and all other training settings held fixed.

For seed 42, weight 5 has the highest AUROC/AUPRC and the lowest CGVR/NZR among the tested values. Increasing the weight to 10 or 20 lowers predictive scores and raises both diagnostic rates relative to weight 5.

Table 5: eICU-derived internal cohort-shift diagnostics on ICU Sepsis under `HADM_ID`-group-level distribution shifts. All results are from seed 42. AUROC and AUPRC measure TSTR association preservation; CGVR and NZR measure directional violations and action collapse, respectively. Action μ/σ gives the mean and population standard deviation of valid generated actions on the shifted test partition. Bold/underline mark the best/second-best displayed scores within each shift; ties share a rank.

Method	AUROC ↑	AUPRC ↑	CGVR (%) ↓	NZR (%) ↓	Action μ/σ	Unique@3dp
A. MAP-burden shift						
TabICLv2	0.6632	0.4217	1.680	1.627	0.0678/0.0426	316
SFT-only LLM†	0.6699	0.4362	87.041	100.000	0.0000/0.0000	1
Weighted SFT	0.6702	0.4329	1.221	39.577	0.0792/0.0661	13
Standard-DPO	<u>0.6751</u>	<u>0.4388</u>	0.656	0.525	0.0920/0.0366	65
SimPO	0.6622	0.4189	55.893	59.722	0.0266/0.0328	18
SafeDPO-CGVR	0.6748	0.4365	0.053	<u>0.075</u>	0.1303/0.0360	108
KTO	0.6700	0.4349	<u>0.043</u>	18.547	0.1025/0.0673	157
VasoDPO	**0.6981**	**0.4590**	**0.003**	**0.005**	0.1096/0.0442	76
B. Severity-risk shift						
TabICLv2	0.4889	0.3252	**0.000**	**0.000**	0.1483/0.0225	135
SFT-only LLM†	<u>0.5164</u>	0.3388	55.220	100.000	0.0000/0.0000	1
Weighted SFT	0.5105	0.3347	1.582	69.428	0.0400/0.0617	10
Standard-DPO	0.5119	<u>0.3412</u>	1.701	0.942	0.1264/0.0303	81
SimPO†	<u>0.5164</u>	0.3388	55.220	100.000	0.0000/0.0000	1
SafeDPO-CGVR	0.5078	0.3347	**0.000**	**0.000**	0.1399/0.0482	165
KTO	0.5126	0.3362	**0.000**	43.127	0.0823/0.0760	112
VasoDPO	**0.5563**	**0.3827**	<u>0.134</u>	<u>0.857</u>	0.0764/0.0440	88

† Zero-action collapse: all generated actions are zero. Unique@3dp counts distinct action values after rounding to three decimals.

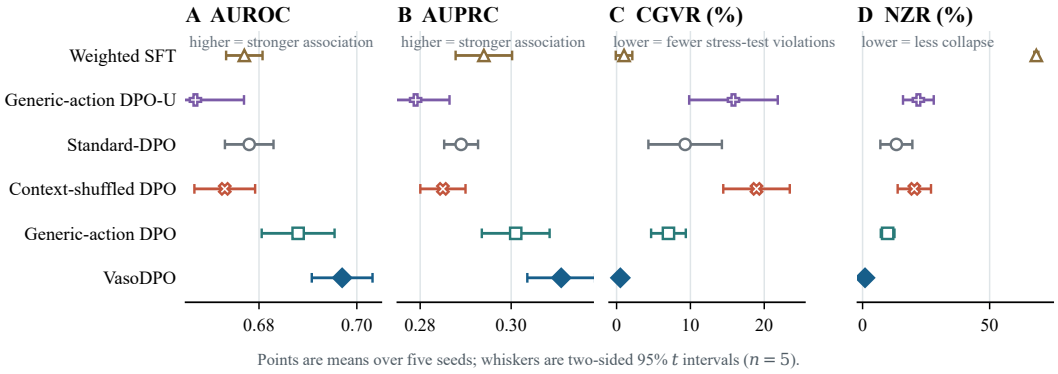

Figure 3: **Five-seed uncertainty for Table 2.** Points are five-seed means; bars are two-sided 95% Student t intervals with four degrees of freedom, computed as specified in Appendix C.6. They summarize optimization-seed variation on the fixed split.

E.6 Threshold and high-action support diagnostics

The following grid separates directional-threshold sensitivity from empirical high-action support.

Evaluator-threshold sensitivity and support. Over the grid defined in Appendix C.5, VasoDPO has mean LMW of 1.035% and mean NZR of 1.267%. SFT-only and Weighted SFT retain high NZR across this grid. Among 7,462 high-MAP test records, 23 observed actions reach 0.70 and three reach the primary threshold of 0.99. None of the displayed generated-policy outputs reaches 0.70 in these states. This sparse upper-tail support explains the limited discrimination of HME at the primary

Table 6: Five-seed generation diagnostics. Performance entries are means ± sample standard deviations over seeds $\{42, 43, 44, 45, 46\}$. Action μ/σ reports the mean of the five within-seed action means and the mean of the five within-seed action standard deviations, respectively. Within-seed action standard deviations use the population denominator. Parser acceptance is reported as a percentage. In Panel A, bold/underline mark the best/second-best mean among learned LLMs, as in Table 1. Panel B reports descriptive summaries.

A. Predictive and action diagnostics

Method	AUROC ↑	AUPRC ↑	CGVR (%) ↓	NZR (%) ↓	LMW (%) ↓
Ridge	0.653 ± 0.000	0.264 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
XGBoost	0.672 ± 0.002	0.284 ± 0.002	1.586 ± 0.046	1.306 ± 0.022	2.908 ± 0.084
TabICLv2	0.667 ± 0.002	0.276 ± 0.002	2.358 ± 0.042	2.200 ± 0.034	4.325 ± 0.077
SFT-only LLM	0.674 ± 0.008	0.295 ± 0.005	53.216 ± 1.210	97.330 ± 2.509	97.589 ± 2.219
Weighted SFT	0.677 ± 0.003	0.294 ± 0.005	1.010 ± 0.904	68.369 ± 0.653	1.852 ± 1.658
Standard-DPO	0.678 ± 0.004	0.289 ± 0.003	9.286 ± 4.009	13.295 ± 5.072	17.030 ± 7.352
Rule-rerank SFT	0.676 ± 0.002	0.292 ± 0.002	34.771 ± 3.716	75.846 ± 6.579	63.764 ± 6.814
SimPO	0.674 ± 0.007	0.294 ± 0.004	45.773 ± 2.825	75.847 ± 2.964	83.940 ± 5.180
KTO	<u>0.686</u> ± 0.004	<u>0.300</u> ± 0.003	1.409 ± 1.001	43.043 ± 4.990	2.583 ± 1.836
SafeDPO-CGVR	0.678 ± 0.002	0.289 ± 0.004	<u>0.874</u> ± 0.722	<u>4.824</u> ± 3.394	<u>1.602</u> ± 1.325
VasoDPO	**0.697** ± 0.005	**0.311** ± 0.006	**0.510** ± 0.033	**0.967** ± 0.013	**0.935** ± 0.061

B. Action summaries

Method	Action μ/σ	Parser acceptance (%) ↑
Ridge	0.0677/0.0172	100.000
XGBoost	0.0681/0.0332	100.000
TabICLv2	0.0678/0.0413	100.000
SFT-only LLM	0.0023/0.0123	100.000
Weighted SFT	0.0436/0.0655	100.000
Standard-DPO	0.0790/0.0444	100.000
Rule-rerank SFT	0.0182/0.0418	100.000
SimPO	0.0153/0.0321	100.000
KTO	0.0761/0.0874	100.000
SafeDPO-CGVR	0.0999/0.0439	100.000
VasoDPO	0.0826/0.0452	100.000

Table 7: **Single-seed omission-weight sensitivity diagnostic.** With $\omega_{\text{std}} = 1$, we vary $\omega_{\text{omit}} \in \{1, 2, 5, 10, 20\}$ on seed 42 while holding the state-fixed preference corpus, optimization settings, and evaluation protocol fixed. The $\omega_{\text{omit}} = 5$ row is the seed-42 VasoDPO run included in the main five-seed aggregate in Table 1. Setting $\omega_{\text{omit}} = 1$ yields uniform pair weights. Bold/underline mark the best/second-best displayed scores per metric; ties share a rank.

ω_{omit}	AUROC ↑	AUPRC ↑	CGVR (%) ↓	NZR (%) ↓	Action μ/σ
1	<u>0.681</u>	0.290	3.516	4.568	0.0927/0.0423
2	<u>0.681</u>	<u>0.299</u>	2.431	4.248	0.0987/0.0426
5	**0.695**	**0.310**	**0.487**	**0.959**	0.0705/0.0201
10	0.671	0.284	<u>1.444</u>	<u>3.282</u>	0.0947/0.0396
20	0.666	0.281	1.621	3.769	0.0960/0.0399

threshold; with HME equal to zero, CGVR reflects the low-MAP component and its share of the eligible records.

E.7 Dose-bin diagnostics

Dose-bin definitions. All diagnostics use normalized norepinephrine actions, with fixed bins near-zero ($a \leq 0.01$), low ($0.01 < a \leq 0.05$), medium ($0.05 < a \leq 0.15$), and high ($a > 0.15$).

Table 8: **Evaluator-threshold sensitivity and high-action support.** Panel A reports LMW and NZR summaries over the prespecified low-MAP/near-zero threshold grid. Mean LMW is the arithmetic mean over the 12 low-MAP/near-zero threshold combinations and five seeds (60 evaluations), whereas mean NZR averages the three distinct near-zero thresholds across five seeds (15 evaluations). Panel B fixes $\tau^{+}_{\mathrm{MAP}} = 0.5$ and reports the fraction of high-MAP-eligible actions at or above $\tau^{1}_{a} \in \{0.30, 0.50, 0.70, 0.99\}$. These are evaluator-threshold sensitivity and high-action support diagnostics for the fixed trained policies. In Panel A, bold/underline mark the best/second-best displayed rates among all listed methods. Panel B reports support frequencies.

A. Low-MAP/near-zero evaluator sensitivity

Method	Mean LMW (%) ↓	Mean NZR (%) ↓
VasoDPO	<u>1.035</u>	<u>1.267</u>
KTO	4.911	43.043
SafeDPO-CGVR	1.679	4.824
SFT-only LLM	97.672	97.330
Ridge	**0.033**	**0.028**
Standard-DPO	18.484	13.295
Weighted SFT	7.330	68.369
XGBoost	4.859	1.767

B. High-MAP high-action support audit

Source / generated policy	$\tau^{1}_{a} = 0.30$	0.50	0.70	0.99 (primary)
Observed test actions	2.814 (210)	0.911 (68)	0.308 (23)	0.040 (3)
VasoDPO	0.051 (19)	0.003 (1)	0.000 (0)	0.000 (0)
KTO	0.005 (2)	0.000 (0)	0.000 (0)	0.000 (0)
SafeDPO-CGVR	0.005 (2)	0.000 (0)	0.000 (0)	0.000 (0)
SFT-only LLM	0.011 (4)	0.005 (2)	0.000 (0)	0.000 (0)
Ridge	0.000 (0)	0.000 (0)	0.000 (0)	0.000 (0)
Standard-DPO	0.024 (9)	0.000 (0)	0.000 (0)	0.000 (0)
Weighted SFT	0.000 (0)	0.000 (0)	0.000 (0)	0.000 (0)
XGBoost	0.188 (70)	0.027 (10)	0.000 (0)	0.000 (0)

Panel B entries are percentages with event counts in parentheses. Observed-action rates use the 7,462 high-MAP-eligible rows in the single held-out test set. Generated-policy rates pool the five seed-specific outputs over the same eligible contexts ($7{,}462 \times 5 = 37{,}310$ seed–row evaluations). Each seed uses the same eligible states, so each pooled percentage equals the mean of the five seed-specific rates. With 100% parser acceptance, it also equals HME at that action threshold.

Bin-level classification scores compare generated and observed bins on held-out rows for which both actions are valid. The evaluator averages class F1 over bins with at least one true positive (TP), labeled F1 (TP>0) in the table. Balanced accuracy averages recall over bins with observed support. Low-MAP medium/high recall is the fraction of valid rows with standardized MAP below -0.5 and an observed medium/high action whose generated action is also medium/high. Metrics are computed per completed seed; Table 9 reports their mean and standard deviation over seeds $\{42, 43, 44, 45, 46\}$.

E.8 TSTR ACTION ABLATION

This ablation compares state-only and state-plus-action downstream predictors under the same split.

These paired differences quantify the change in held-out mortality prediction when generated actions are added to the training features under the action replacement protocol.

E.9 ACTION DIAGNOSTIC LANDSCAPE

Figure 4 provides an auxiliary two-dimensional view of the CGVR and NZR summaries from Tables 1 and 2. SFT-only, Rule-rerank SFT, and SimPO have high values on both displayed diagnostics; Weighted SFT and KTO pair low CGVR with substantial near-zero collapse; and tabular controls occupy the lower-left region. VasoDPO has the lowest displayed mean CGVR and NZR among the learned LLM policies.

Table 9: Dose-bin diagnostics (five-seed means ± standard deviations). Panel B reports NZR and medium/high dose percentages; the omitted low-dose share completes 100% per seed. Bold/underline: best/second-best displayed mean across all methods for Panel A and NZR.

A. Agreement with recorded dose bins

Method	F1 (TP> 0) ↑	Balanced accuracy ↑	Low-MAP medium/high recall (%) ↑
Ridge	**0.364** ± 0.000	0.268 ± 0.000	84.957 ± 0.000
XGBoost	0.216 ± 0.001	0.284 ± 0.001	73.376 ± 0.319
TabICLv2	0.217 ± 0.001	0.280 ± 0.001	63.319 ± 0.316
SFT-only LLM	0.302 ± 0.203	0.252 ± 0.003	2.959 ± 2.922
Weighted SFT	0.256 ± 0.005	0.224 ± 0.003	**99.403** ± 0.498
Standard-DPO	0.266 ± 0.020	<u>0.292</u> ± 0.013	82.895 ± 9.352
Rule-rerank SFT	0.217 ± 0.013	0.269 ± 0.002	28.829 ± 4.963
SimPO	0.229 ± 0.011	0.284 ± 0.008	10.979 ± 8.237
KTO	0.245 ± 0.008	0.264 ± 0.010	97.766 ± 1.275
SafeDPO-CGVR	0.225 ± 0.019	0.276 ± 0.007	98.397 ± 1.364
VasoDPO	<u>0.312</u> ± 0.018	**0.303** ± 0.012	<u>98.525</u> ± 0.609

B. Generated action distribution

Method	NZR (%) ↓	Medium (%)	High (%)
Ridge	**0.000** ± 0.000	85.582 ± 0.000	0.000 ± 0.000
XGBoost	1.306 ± 0.022	72.618 ± 0.354	2.344 ± 0.087
TabICLv2	2.200 ± 0.034	59.427 ± 0.104	4.382 ± 0.016
SFT-only LLM	97.330 ± 2.509	1.990 ± 1.888	0.149 ± 0.158
Weighted SFT	68.369 ± 0.653	27.167 ± 0.361	4.465 ± 0.485
Standard-DPO	13.295 ± 5.072	71.744 ± 1.328	5.303 ± 2.035
Rule-rerank SFT	75.846 ± 6.579	13.604 ± 4.079	0.929 ± 0.616
SimPO	75.847 ± 2.964	12.383 ± 5.631	0.626 ± 0.632
KTO	43.043 ± 4.990	19.017 ± 3.681	23.473 ± 6.334
SafeDPO-CGVR	4.824 ± 3.394	77.478 ± 6.904	13.671 ± 5.315
VasoDPO	<u>0.967</u> ± 0.013	79.237 ± 3.386	6.815 ± 1.792

Table 10: TSTR action ablation (five-seed means ± standard deviations). State-only AUROC/AUPRC are 0.675 ± 0.000/0.294 ± 0.000 for every method. Action-augmented predictors use generated training actions and recorded test actions. Each Δ subtracts the state-only score before rounding. Bold/underline: best/second-best displayed mean across methods.

	AUROC ↑		AUPRC ↑	
Method	With action	Δ ↑	With action	Δ ↑
Ridge	0.653 ± 0.000	−0.022 ± 0.000	0.264 ± 0.000	−0.029 ± 0.000
XGBoost	0.672 ± 0.002	−0.003 ± 0.002	0.284 ± 0.002	−0.010 ± 0.002
TabICLv2	0.667 ± 0.002	−0.008 ± 0.002	0.276 ± 0.002	−0.017 ± 0.002
SFT-only LLM	0.674 ± 0.008	−0.001 ± 0.008	0.295 ± 0.005	0.002 ± 0.005
Weighted SFT	0.677 ± 0.003	0.002 ± 0.003	0.294 ± 0.005	0.000 ± 0.005
Standard-DPO	0.678 ± 0.004	0.003 ± 0.004	0.289 ± 0.003	−0.005 ± 0.003
Rule-rerank SFT	0.676 ± 0.002	0.001 ± 0.002	0.292 ± 0.002	−0.001 ± 0.001
SimPO	0.674 ± 0.007	−0.001 ± 0.007	0.294 ± 0.004	0.000 ± 0.004
KTO	<u>0.686</u> ± 0.004	<u>0.011</u> ± 0.004	<u>0.300</u> ± 0.003	<u>0.007</u> ± 0.003
SafeDPO-CGVR	0.678 ± 0.002	0.003 ± 0.002	0.289 ± 0.004	−0.005 ± 0.004
VasoDPO	**0.697** ± 0.005	**0.022** ± 0.005	**0.311** ± 0.006	**0.017** ± 0.006

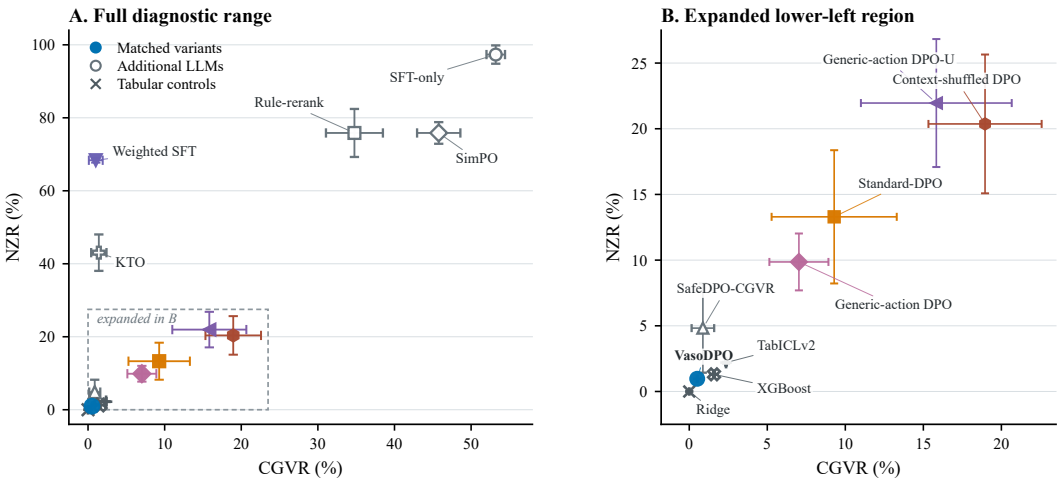

Figure 4: **Action diagnostic landscape across learned policies and controls.** Panel A shows the full range; Panel B magnifies the dashed lower-left region. Filled markers denote matched variants, open markers denote other LLM baselines, and crosses denote tabular controls; whiskers are seed standard deviations where applicable. Lower coordinates indicate fewer displayed directional violations and less near-zero collapse under the prespecified thresholds.