MEDROAM: IMPROVING ROLLOUT ROBUSTNESS AND INTERVENTION SENSITIVITY IN MEDICAL LANGUAGE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical language world models (LWMs) use large language model (LLM) backbones to predict how a patient's state evolves given the clinical history and an intervention. Despite their promise, medical LWMs remain at an early stage of development, and their limitations are not yet well understood. Through a series of preliminary analyses, we identify two key shortcomings: (1) *limited rollout robustness*, reflected in a substantial performance drop from teacher-forced to autoregressive simulation, and (2) *weak intervention sensitivity*, where predictions barely change even when the probing intervention is removed. To address these shortcomings, we propose MEDROAM, which jointly optimizes two simple-yet-effective training objectives: *Privileged Self-Distillation* (PSD) and *Intervention Response Alignment* (IRA). PSD uses a frozen teacher conditioned on ground-truth histories to guide a student trained on a mixture of clean and noisy rollout histories, encouraging robustness to accumulated errors while anchoring predictions to the initial model. IRA incorporates pharmacological knowledge to align numerical response estimates across the original intervention, its removal, and a similarly acting substitute. Extensive experiments across several LLM backbones show that MEDROAM improves autoregressive rollout accuracy on in-distribution and out-of-distribution data while largely preserving teacher-forced performance. Evaluations on 14 supervised and 7 held-out intervention-response probing pairs further demonstrate improved intervention sensitivity and pharmacological consistency.

1 INTRODUCTION

World Models (WMs), which aim to learn the dynamics of an environment, have attracted growing attention in recent years (Ha & Schmidhuber, 2018; LeCun, 2022; Xiang et al., 2025; Ding et al., 2025). Given a history of states and an action, they predict how the environment will evolve, allowing an agent to roll out and compare the consequences of candidate actions in imagination, without executing them in the real world (Hafner et al., 2019). Clinical decision-making is a natural fit for this paradigm (Qazi et al., 2025). Conventional clinical prediction models are static and observational, *e.g.*, estimating the risk of mortality or of a specific complication from a patient's current state. These models can answer *what will happen to this patient*, but not *what would happen under a different intervention*, which is the question that actually drives clinical decisions (Feng et al., 2026). This gap motivates the study of medical WMs (Saeed et al., 2026; Yang et al., 2025b).

Several existing medical WMs encode patient states as latent representations and learn transitions in the resulting latent space (Xu et al., 2026b; Ding et al., 2026; Wu et al., 2026). These compact representations enable efficient simulation, but may limit the preservation of detailed information from long clinical histories. Meanwhile, limited intervention coverage in observational data poses challenges for extrapolation to unfamiliar interventions. Moreover, most medical WMs focus primarily on structured measurements, while neglecting the textual information that is also informative in heterogeneous electronic health records (EHRs) (Johnson et al., 2023). More recently, **medical language world models** (LWMs) have emerged as an alternative, using large language models (LLMs) (Minaee et al., 2024; Zhao et al., 2026b) to predict patient states from clinical histories and candidate interventions. Their pretrained medical knowledge and text-processing capabilities may facilitate clinical trajectory modeling with sparse data and support extrapolation to out-of-distribution (OOD) interventions.

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.

Despite this potential, research on medical LWMs remains limited. A representative early effort is EHRWorld (Mu et al., 2026), which trains a series of Qwen3 models (Yang et al., 2025a) on a large-scale real-world EHR dataset, MIMIC-IV (Johnson et al., 2023). Given a patient's history and current intervention, EHRWorld predicts the values of subsequent clinical channels, thereby modeling clinical trajectories as sequences of intervention-conditioned observations.

To assess the capabilities and limitations of medical LWMs, we take EHRWorld as a representative case and conduct experiments along two key axes: *rollout robustness* and *intervention sensitivity*. Both are essential for counterfactual reasoning, *i.e.*, an LWM must maintain accuracy when rolling out over its own predictions and respond meaningfully to the specified interventions. Our experiments reveal two limitations: ❶ a substantial performance gap between teacher-forced and autoregressive simulation; and ❷ only minor changes in model predictions even when critical probing interventions are removed. These findings raise concerns about the reliability of simulated trajectories and their utility for clinical decision-making, motivating the question: *can we train medical language world models that are both robust over long horizons and sensitive to interventions?*

In this paper, we propose MEDROAM, a framework for **Ro**bust and intervention-**A**ware medical language world **M**odeling, which addresses these limitations via two simple training objectives. First, to address ❶, *Privileged Self-Distillation* (PSD) provides a frozen teacher with observed ground-truth histories as privileged information to guide a student initialized from the same LWM. Unlike costly on-policy distillation (Agarwal et al., 2024), PSD reuses a fixed set of offline autoregressive rollouts and trains the student on a mixture of ground-truth and cached noisy rollout histories, promoting rollout robustness while preserving teacher-forced accuracy. Second, to address ❷, *Intervention Response Alignment* (IRA) uses curated intervention–response pairs with pharmacological priors to constrain differentiable numerical response estimates. For each pair, we compare predictions under the original intervention, its removal, and substitution with a similarly acting intervention under the same observed history. The alignment loss encourages pharmacologically consistent differences between the original and removal conditions, while limiting discrepancies under substitution. In general, we jointly optimize PSD and IRA in a single refinement stage to improve rollout robustness and the pharmacological consistency of intervention-conditioned predictions.

Empirically, following EHRWorld, we construct a longitudinal clinical dataset from real-world ICU EHRs (Johnson et al., 2023) and train LWMs with Qwen3-4B/8B/14B backbones. We mainly evaluate these models along two dimensions. First, we evaluate rollout accuracy under teacher-forced and autoregressive settings on both in-distribution (ID) and OOD test data. Then, we further develop four-level intervention probing tasks, which assess removal responsiveness (L1), directional consistency (L2), dose sensitivity (L3), and intervention specificity (L4). Extensive results show that MEDROAM improves autoregressive rollout accuracy while largely preserving teacher-forced performance, and improves intervention responsiveness and agreement with pharmacological expectations. To summarize, our contributions are threefold: (1) We identify and characterize two key limitations of medical LWMs: limited rollout robustness and weak intervention sensitivity. (2) We propose MEDROAM, which jointly optimizes PSD and IRA to address these limitations. (3) We develop a four-level intervention probing framework, and demonstrate the effectiveness of MEDROAM through ID and OOD rollout evaluations and intervention probing across model architectures and scales.

2 DIAGNOSING MEDICAL LANGUAGE WORLD MODELS

2.1 PRELIMINARIES

Clinical Trajectory Representation. We represent a clinical trajectory in EHRs as a sequence of intervention-to-observation transitions. At transition t, the intervention a_t specifies the medication, dosage, and route of administration. The state $s_t = \{(q, v_{t,q}) \mid q \in \mathcal{Q}_t\}$ is the post-intervention prediction target, where $\mathcal{Q}_t$ denotes the target clinical channels (*e.g.*, laboratory tests and vital signs), and $v_{t,q}$ is the corresponding observed value. Notably, $\mathcal{Q}_t$ varies across transitions, reflecting irregular sampling and missingness in EHRs. Let c denote the static clinical context, *i.e.*, basic patient information (age, gender, and ICU types). We can define the trajectory as $\tau = \big(c, (a_t, s_t)_{t=1}^{T}\big)$, where T is the number of transitions. The history $h_t = \{(a_i, s_i)\}_{i<t}$ is the ordered event prefix before a_t.

Medical Language World Model. A medical LWM $\mathcal{M}_\theta$ predicts the post-intervention target s_t conditioned on c, h_t, and a_t. Let $\mathrm{Serialize}(s_t) = (y_{t,1}, \ldots, y_{t,L_t})$ denote the serialized token

sequence of s_t, where L_t is its length, and let $\mathcal{V}_t \subseteq \{1, \ldots, L_t\}$ index the tokens corresponding to target channel values. Only these value tokens contribute to the prediction loss, while the other tokens (*e.g.*, channel names, timestamps, and units) are excluded. By doing so, we can denote the likelihood of the serialized value tokens by $p_\theta(s_t \mid c, h_t, a_t)$ and define it as follows:

$$p_\theta\left(s_t \mid c, h_t, a_t\right) = \prod_{l \in \mathcal{V}_t} p_\theta\left(y_{t,l} \mid y_{t,<l},\, c,\, h_t,\, a_t\right). \tag{1}$$

The medical LWM is trained with the standard maximum likelihood objective on target value tokens:

$$\mathcal{L}_{\text{LWM}} = -\sum_{t=1}^{T} \sum_{l \in \mathcal{V}_t} \log p_\theta\left(y_{t,l} \mid y_{t,<l},\, c,\, h_t,\, a_t\right). \tag{2}$$

Autoregressive Rollout. Given a sequence of candidate interventions $(a_t, \ldots, a_{t+N-1})$, the medical LWM simulates the patient trajectory over N future transitions. Starting from the observed history $\hat{h}_t = h_t$, each predicted state $\hat{s}_{t+i} \sim p_\theta(\cdot \mid c, \hat{h}_{t+i},\, a_{t+i})$ is fed back into the history through $\hat{h}_{t+i+1} = \hat{h}_{t+i} \oplus (a_{t+i}, \hat{s}_{t+i})$ for $i \in \{0, \ldots, N-1\}$, where $\oplus$ denotes chronological appending.

2.2 Empirical Analyses

Settings. To obtain the medical LWMs, we fine-tune Qwen3-4B/8B/14B models on real-world ICU EHR data from MIMIC-IV. Specifically, following the data curation in EHRWorld, we mainly extract basic patient information (age, gender, and ICU type) from corresponding clinical tables for each ICU stay (Appendix B.1), and serialize the patient history as interleaved inquiry and intervention events, with s_t denoting the target inquiries following a_t. Then, we use the same filtering criteria as EHRWorld to ensure data quality and consistency. As a result, we obtain 54,551 longitudinal clinical trajectories and split them into 51,624 for training, 200 for dev, and 2,727 for testing. For model training, we use the same hyperparameters as EHRWorld, and the details are shown in §4.1.

Findings. Through extensive analyses, we find that current medical LWMs fall short in two fundamental capabilities: *rollout robustness* and *intervention sensitivity*. Specifically,

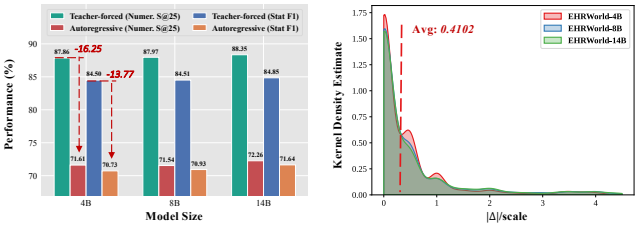


Figure 1: **Left.** Teacher-forced and autoregressive rollout performance across model sizes. **Right.** Distributions of absolute prediction changes after probing intervention removal.

❶ Limited Rollout Robustness: Since accurate simulation of future patient states is essential for medical WMs, we first compare state prediction performance under teacher-forced and autoregressive rollout, where the former conditions on ground-truth histories and the latter feeds the model's own predictions back into the history. Following EHRWorld, we adopt *Numer. S@25* and *Stat F1* as evaluation metrics (Appendix C.3). As shown in Figure 1 (**Left**), all models perform well under teacher forcing, reaching up to 88.35 in *Numer. S@25* and 84.85 in *Stat F1*. However, in the autoregressive rollout, the performance of all models degrades significantly. In the case of EHRWorld-4B, the performance drops are 16.25 and 13.77 points in *Numer. S@25* and *Stat F1*, respectively. One of the main reasons is the training-inference mismatch, as early errors accumulate and corrupt the subsequent predictions. **Takeaway**: *a substantial gap remains between teacher-forced and autoregressive rollout in medical LWMs.*

❷ Weak Intervention Sensitivity. Using 14 curated intervention-response pairs (see Appendix B.4), we assess intervention sensitivity by removing the probing intervention a_t and measuring changes in the corresponding predicted channel values in s_t. All comparisons keep the recorded history fixed and remove only the probing intervention at the current step. To quantify sensitivity, we compute the absolute prediction difference Δ between the original and removal conditions and normalize it by the channel-specific scale, defined as the median absolute difference between consecutive measurements. As shown in Figure 1 (**Right**), the normalized difference is zero for most samples, and its sample-level mean is only 0.4102 for EHRWorld-4B. These results suggest that model predictions often remain unchanged even when an intervention relevant to the target channel is removed. **Takeaway:** *only minor changes in model predictions occur even when critical probing interventions are removed.*

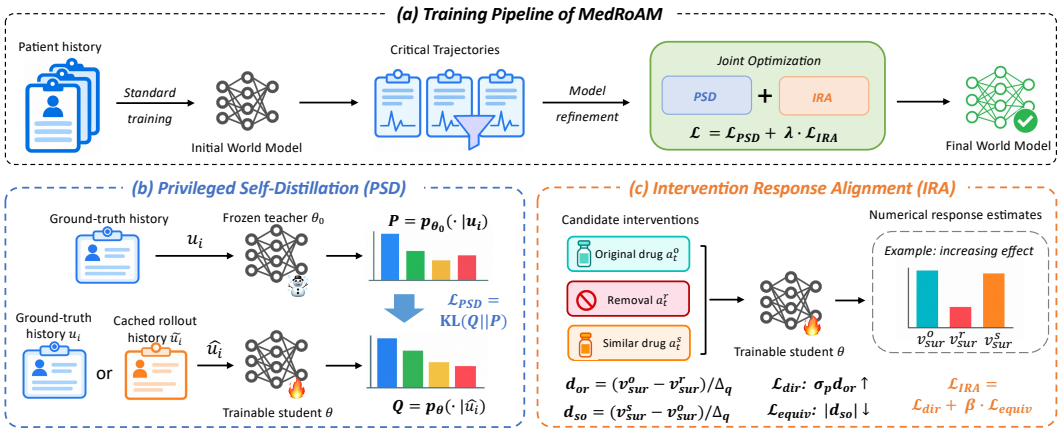


Figure 2: **Overview of MEDROAM.** (**a**) We refine an initially trained LWM with PSD and IRA on critical trajectories containing curated intervention-response pairs. (**b**) PSD distills a frozen teacher conditioned on ground-truth histories into a student using clean or noisy rollout histories. (**c**) IRA uses ground-truth histories to align numerical responses across original, removed, and similarly substituted interventions, promoting directional alignment and substitution consistency.

3 METHODOLOGY

3.1 MOTIVATION AND INTUITION OF MEDROAM

To address these limitations, we propose MEDROAM, consisting of two training strategies. On the one hand, our preliminary analyses show that medical LWMs predict more accurately under teacher-forcing than autoregressive rollout. This suggests that predictions conditioned on observed clean histories can provide useful supervision for learning from imperfect histories. Inspired by on-policy self-distillation (Zhao et al., 2026a), we introduce **Privileged Self-Distillation (PSD)**, which uses a frozen teacher with observed past states as privileged information to guide a student conditioned on self-generated histories. This encourages robustness to accumulated errors and narrows the training-inference mismatch. On the other hand, supervision on target value tokens alone does not specify how predictions should respond to intervention changes, motivating an additional learning signal. To provide this signal, we curate 14 representative intervention-response pairs from the training data with pharmacologically supported effect directions (Appendix B.4). Using these pairs, **Intervention Response Alignment (IRA)** aligns numerical response changes with pharmacological directions and encourages consistency under similarly acting substitutions. Figure 2 provides an overview.

3.2 PRIVILEGED SELF-DISTILLATION

Training data construction. MEDROAM aims to refine an existing medical LWM $\mathcal{M}_{\theta_0}$ trained with Eq. 2. For training, we select trajectories that support both PSD and IRA. Specifically, we retain trajectories containing at least one critical intervention with an eligible response matching one of the 14 curated pairs (Appendix B.4), ensuring that each trajectory supports pharmacological alignment. Since repeated on-policy sampling is costly, we generate autoregressive rollouts offline using $\mathcal{M}_{\theta_0}$ on this subset. Each predicted value is fed back into the history, while the clinical context, recorded interventions, and measurement schedule remain fixed. Moreover, we recompute abnormality tags from generated values to keep histories internally consistent. The generated trajectories are paired with the originals, and are cached and reused throughout optimization without further sampling.

Self-distillation training. We freeze the initial LWM $\mathcal{M}_{\theta_0}$ as the teacher and initialize the student $\mathcal{M}_\theta$ with $\theta = \theta_0$. For paired inquiry i, let u_i and $\tilde{u}_i$ denote the ground-truth and noisy rollout prefixes ending immediately before their response value. Both include the current inquiry but exclude its value and future observations. The teacher always receives u_i, *i.e.*, using the observed past states as its privileged information. Although training on rollout histories encourages robustness to accumulated errors, it may fail to preserve predictions on recorded histories. Thus, we sample the student's input history per example, choosing the recorded history with probability ρ and the cached rollout with probability $1 - \rho$. All inquiries within an example share this choice, with the selected prefix denoted

by $\hat{u}_i \in \{u_i, \tilde{u}_i\}$. The recorded branch anchors the student to the initial model on its original inputs, while the rollout branch teaches it to recover teacher predictions from imperfect histories.

Since generated and ground-truth values may differ in token length, we pair inquiry occurrences in the shared serialized trajectory, adjusting rollout positions for changes in preceding value lengths. Thus, repeated measurements of the same channel remain distinct. We align next-token distributions only at the first token of each value, before either model consumes part of the current number. This avoids comparing later tokens conditioned on different partial numbers. For all paired inquiries $\mathcal{I}$ in a batch, we can define the training objective of PSD as:

$$\mathcal{L}_{\mathrm{PSD}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathrm{KL}(p_\theta(\cdot \mid \hat{u}_i) \,\|\, p_{\theta_0}(\cdot \mid u_i)) \,, \tag{3}$$

where reverse KL is computed over the full vocabulary and gradients update only the student. Notably, we use reverse KL for its stronger empirical performance, with alternatives compared in Appendix D.1. The student processes only the sampled history, avoiding separate forward passes for both histories.

3.3 Intervention Response Alignment

Candidate interventions. At each eligible critical step t, we construct three arms: the original intervention $a_t^{\mathrm{o}} = a_t$, its removal $a_t^{\mathrm{r}} = \mathrm{Remove}(a_t)$, and a similarly acting substitute $a_t^{\mathrm{s}} = \mathrm{Sub}(a_t)$. Removal provides a reference for the predicted contribution of a_t, while substitution encourages consistency across similarly acting drugs. Across all three arms, we edit only the target intervention and keep the remaining history and response query fixed. Substitutes follow Appendix B.5 (Table 8), prioritizing the same drug class and using their most frequent training-set order strings.

Alignment training. To align numerical responses through gradient-based training, we represent $v_{t,q}$ with a differentiable surrogate without discrete decoding. Let $\mathcal{T}_{\mathrm{num}}$ denote the set of tokens representing nonempty strings of ASCII decimal digits, and let $n(u)$ be the integer value represented by token u. Under arm $b \in \{\mathrm{o}, \mathrm{r}, \mathrm{s}\}$, let $P_k^b(u)$ denote the full-vocabulary softmax probability of token u at numeric-token position k, conditioned on the corresponding intervention and the teacher-forced prefix of the observed response. We compute $v_{\mathrm{sur}}^b = \eta \sum_{k=1}^{K} \pi_k \sum_{u \in \mathcal{T}_{\mathrm{num}}} n(u) P_k^b(u)$, where the sign η, the number of numeric tokens K, and their place values π_k are fixed by the tokenization of the observed response and shared across arms. Each π_k is the decimal place value of the rightmost digit of the corresponding observed token (more cases in different tokenizers are in Appendix C.1). This construction accommodates tokens containing one or multiple digits and reduces to digit-wise weighting for single-digit tokenizers. It provides a differentiable numerical surrogate under a shared observed-response format. Since channels differ in units and variability, we normalize differences by Δ_q, the median absolute difference between consecutive measurements of channel q in the training split, floored at 10^{-6} to prevent division by zero. We define $d_{\mathrm{or}} = (v_{\mathrm{sur}}^{\mathrm{o}} - v_{\mathrm{sur}}^{\mathrm{r}})/\Delta_q$ for the original intervention relative to removal and $d_{\mathrm{so}} = (v_{\mathrm{sur}}^{\mathrm{s}} - v_{\mathrm{sur}}^{\mathrm{o}})/\Delta_q$ for substitution relative to the original. For pair p, the supported direction is $\sigma_p = +1$ for an increase and -1 for a decrease. The two instance-level alignment losses in IRA can be defined as follows:

$$\mathcal{L}_{\mathrm{dir}} = [m - \sigma_p d_{\mathrm{or}}]_+, \qquad \mathcal{L}_{\mathrm{equiv}} = \big[|d_{\mathrm{so}}| - \gamma \max\{\mathrm{sg}[|d_{\mathrm{or}}|], m\}\big]_+, \tag{4}$$

where $[x]_+ = \max(0, x)$, $m > 0$ is a margin, and $0 < \gamma < 1$ controls substitution tolerance. The directional loss $\mathcal{L}_{\mathrm{dir}}$ encourages a separation of at least m in the supported direction, discouraging negligible or reversed effects. The substitution loss $\mathcal{L}_{\mathrm{equiv}}$ permits limited variation because similarly acting drugs can differ in potency. Its tolerance scales with $|d_{\mathrm{or}}|$ and has a floor of γm, avoiding a requirement for nearly identical responses when the predicted effect is weak. The stop-gradient operator sg preserves its input but blocks gradients through it, preventing $\mathcal{L}_{\mathrm{equiv}}$ from directly increasing the original–removal difference to enlarge its own tolerance. Notably, to avoid combining pharmacological alignment with accumulated rollout errors, we apply IRA only to recorded-history examples, with all three arms sharing the same observed history. To bound the loss contribution of instances with extreme directional violations, we cap the directional hinge loss as $\overline{\mathcal{L}}_{\mathrm{dir}} = \min\{\mathcal{L}_{\mathrm{dir}}, 2m\}$. When $\sigma_p d_{\mathrm{or}} < -m$, the loss saturates at $2m$ and contributes no directional gradient.

3.4 Joint Training

We optimize PSD and IRA jointly on the selected subset. Specifically, PSD supervises all paired inquiries under sampled student histories, while IRA applies only to the recorded-history branch.

Thus, ρ controls how frequently examples receive IRA supervision. Specifically, let $\mathcal{E}_{\text{obs}}$ denote the recorded-history examples in the current microbatch. Each eligible instance contributes $\mathcal{L}_{\text{IRA}} = \mathcal{L}_{\text{dir}} + \beta\mathcal{L}_{\text{equiv}}$, where β controls substitution consistency. We obtain $\mathcal{L}_{\text{IRA}}$ by averaging these losses first within each example and then across $\mathcal{E}_{\text{obs}}$. If no eligible instances are present, including when $\mathcal{E}_{\text{obs}} = \varnothing$, we set $\mathcal{L}_{\text{IRA}} = 0$. Overall, the joint training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{PSD}} + \lambda\mathcal{L}_{\text{IRA}}, \tag{5}$$

where λ controls the overall alignment strength. Notably, only the student is updated, and the teacher and cached rollouts remain fixed throughout MEDROAM training.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We follow the EHRWorld data processing pipeline to construct a large-scale longitudinal clinical dataset from real-world ICU EHRs in MIMIC-IV (Johnson et al., 2023). The dataset comprises 54,551 longitudinal clinical trajectories, with 51,624 used for initial model training, 200 for model development, and 2,727 reserved for in-distribution (ID) evaluation. Following the criteria in Appendices B.3-B.4, we identify 5,646 critical trajectories in the training set, each containing at least one critical probing intervention, to form the training corpus for MEDROAM. For rollout evaluation, in addition to the ID test data, we also evaluate out-of-distribution (OOD) generalization on eICU-CRD (Demo) (Pollard et al., 2018), an external multi-center ICU dataset, containing 1,333 test trajectories. For intervention evaluation, we select 14 representative intervention-response pairs used for IRA supervision, denoted as held-in probes, and 7 additional pairs excluded from IRA supervision for held-out probing. Further details of data curation are provided in Appendix B.

Evaluation Metrics. For rollout evaluation, we follow EHRWorld and report *Numer. S@25* and *Stat F1* under teacher-forced ("Next") and autoregressive ("Full") settings. Both use the same recorded intervention sequence and measurement schedule, including query times and channels. In "Full", the model predicts all inquiry values without ground-truth inquiry values and recursively feeds its predictions into the history. Interventions and measurement events are provided rather than generated. We also update status tags (*e.g.*, normal or abnormal) to match predicted values, preventing recorded-status leakage. Additional rollout settings are evaluated in Appendix D.3. Moreover, we report $\text{Ret} = (\text{Full}/\text{Next}) \times 100\%$, measuring the proportion of teacher-forced performance retained during autoregressive rollout. For intervention evaluation, we construct four levels of probing tasks (L1–L4) using the curated intervention-response pairs (Appendix B.5). For each probe, we search the full test set and construct one instance per trajectory containing the target drug, using its first administration. We use the ground-truth history and vary only the target intervention, keeping the remaining context, target channel, and query time fixed. Each instance queries one numerical response at the geometric midpoint of its pharmacological window. We define the L1–L4 metrics as follows.

- **L1: Removal Responsiveness.** We remove the target intervention a_t from the context and report *Sens*, the percentage of instances whose target response changes.
- **L2: Directional Consistency.** Among instances whose target responses change after removing a_t, we measure the percentage, *Mono*, that exhibit changes in the pharmacologically expected direction. We also report $\text{Score} = \text{Sens} \times \text{Mono}/100$ to reflect the overall performance, computed from percentage-valued *Sens* and *Mono* within each probe and then macro-averaged across probes.
- **L3: Dose Sensitivity.** We vary the dose of a_t and measure the median Spearman rank correlation, denoted as *Dose-*ρ, between dose and the predicted target response, with signs adjusted so that positive correlations indicate alignment with the pharmacologically expected direction.
- **L4: Intervention Specificity.** First, we remove an unrelated intervention and report *USR*, the ratio of macro-averaged control to positive-probe sensitivities. Lower USR indicates lower relative, but not necessarily absolute, control sensitivity. Next, we replace a_t with a pharmacologically similar intervention and report *SER*, the percentage showing a nonzero substitution effect smaller than the removal effect, among instances with valid predictions in all three arms and a nonzero removal effect. Finally, we use an opposing intervention and report *Rev*, the proportion whose response shifts in the expected opposite direction relative to the prediction under the original intervention.

Table 1: **State prediction performance** (%) under teacher-forced ("Next") and autoregressive ("Full") rollouts on ID and OOD data. The best and second-best results are in **bold** and underlined.

Model	Numer. S@25 (ID)			Stat F1 (ID)			Numer. S@25 (OOD)			Stat F1 (OOD)		
	Next	Full	Ret.	Next	Full	Ret.	Next	Full	Ret.	Next	Full	Ret.
(a) Cutting-edge Large Language Models												
Qwen3-4B	82.14	61.59	74.98	79.30	61.11	77.06	76.97	57.46	74.65	76.37	63.53	83.19
Qwen3-8B	83.57	67.13	80.33	80.79	63.42	78.50	79.82	65.25	81.75	79.22	67.68	85.43
Qwen3-14B	83.37	65.74	78.85	80.77	62.77	77.71	79.45	63.51	79.94	78.80	65.19	82.73
Qwen3.8-27B	84.61	66.11	78.13	82.46	63.99	77.60	81.39	64.02	78.66	81.37	67.68	83.18
DeepSeek-V4-Flash	82.32	67.57	82.08	81.16	65.35	80.52	76.21	64.61	**84.78**	77.35	66.59	86.09
MedGemma-27B	83.29	67.63	81.20	81.15	67.02	82.59	79.20	65.77	83.04	79.53	70.34	88.44
(b) Latent World Models												
CWM	85.55	65.07	76.06	78.19	61.14	78.19	78.70	29.30	37.23	78.10	52.90	67.73
ODE-RNN	85.93	67.93	79.05	78.61	62.13	79.04	78.50	49.40	62.93	78.20	61.40	78.52
(c) Medical Language World Models												
EHRWorld-4B	87.86	71.61	81.50	84.50	70.73	83.70	81.34	65.08	80.01	81.37	67.12	82.49
EHRWorld-8B	87.97	71.54	81.32	84.51	70.93	83.93	81.44	65.24	80.11	81.43	68.74	84.42
EHRWorld-14B	**88.35**	72.26	81.79	**84.85**	71.64	84.43	**82.05**	66.32	80.83	81.68	69.49	85.08
EHRWorld-4B (align)	87.90	71.75	81.63	84.57	71.07	84.04	81.35	64.84	79.70	81.35	67.65	83.16
EHRWorld-8B (align)	87.94	71.54	81.35	84.54	71.25	84.28	81.52	65.08	79.83	81.53	68.33	83.81
EHRWorld-14B (align)	88.30	72.27	81.85	84.83	71.72	84.55	82.00	66.24	80.78	81.65	69.40	85.00
MEDROAM-4B (Ours)	87.49	73.36	**83.85**	84.06	74.60	88.75	80.66	66.42	82.35	81.11	72.68	89.61
MEDROAM-8B (Ours)	87.86	73.37	83.51	84.13	74.80	88.91	80.67	66.27	82.15	81.09	72.33	89.20
MEDROAM-14B (Ours)	88.31	**73.87**	83.65	84.61	**75.25**	**88.94**	81.43	**66.97**	82.24	**81.72**	**73.84**	**90.36**

Implementation Details. We mainly apply MEDROAM to Qwen3 family models, and the results on more LLM backbones are shown in Appendix D. Specifically, following EHRWorld, we first fine-tune the Qwen3-4B/8B/14B instruction models on the full training corpus for three epochs with a learning rate of 1e-6, obtaining the initial model $\mathcal{M}_{\theta_0}$. Then, we apply MEDROAM to further fine-tune the student $\mathcal{M}_\theta$ initialized from this checkpoint on the selected training subset for one epoch with a learning rate of 1e-7. In MEDROAM, λ and β are set to 0.1 and 0.3, respectively. For PSD, we set the recorded-history sampling probability to $\rho = 0.5$ and use the reverse KL divergence. For IRA, we set the margin m and tolerance γ to 2.0 and 0.5, respectively, and cap the loss $\mathcal{L}_{\text{dir}}$ at 4.0 for stable training. At inference time, we use greedy decoding for reproducibility.

Baselines. We compare our models against a diverse set of baselines: (1) **Cutting-edge LLMs**, including five general-purpose LLMs, *Qwen3-4B/8B/14B* (Yang et al., 2025a), *Qwen3.8-27B* (Qwen Team, 2026), *DeepSeek-V4-Flash* (Xu et al., 2026a), and a medical LLM, *MedGemma-27B* (Sellergren et al., 2025). All LLM baselines are evaluated using the same continuation protocol as the LWMs. (2) **Latent WMs**, which encode patient states into a latent space and learn state transitions within it, represented by *CWM* (Wu et al., 2026) and *ODE-RNN* (Rubanova et al., 2019). (3) **Medical LWMs**, the closest counterpart to our work. Since few LWMs have been developed for this setting, we compare mainly against *EHRWorld* (Mu et al., 2026). For fairness, we also continue fine-tuning the EHRWorld models on the same selected training subset and denote them as *EHRWorld (align)*. The implementation and evaluation details of these baselines are provided in Appendix C.2.

4.2 ROLLOUT ROBUSTNESS

For rollout evaluation, we compare rollout accuracy of different models under teacher-forced and autoregressive settings on both ID and OOD test sets. Table 1 presents the detailed comparative results, from which we can make the following key observations.

Obs. 1. Existing WMs improve trajectory prediction but remain limited in OOD generalization and rollout robustness. Despite their smaller size, latent WMs remain competitive with leading LLMs on ID test data but underperform them on most OOD metrics, suggesting limited generalization beyond the observed training trajectories. Conversely, by building on pretrained LLMs, EHRWorld benefits from broader prior medical knowledge and improves over the Qwen3 baselines by an average of 8.67 points in *Stat F1* (ID, Full). Despite this, a substantial performance gap between teacher-forced and autoregressive rollout remains, with ID *Stat F1* retention rates of only 83.70%, 83.93%, and 84.43% for EHRWorld-4B/8B/14B, respectively. These results suggest that language world modeling improves trajectory prediction but still suffers from the training-inference mismatch.

Table 2: **Intervention probing results across L1–L4 tasks.** These tasks assess removal responsiveness (L1), directional consistency (L2), dose sensitivity (L3), and intervention specificity (L4). Notably, "H-In" and "H-Out" denote held-in and held-out intervention–response pairs, respectively.

Method	*Intervention Removal* (L1–L2)						*Dose* (L3)		*Intervention Specificity* (L4)					
	Sens (%, ↑)		Mono (%, ↑)		Score (%, ↑)		Dose-ρ (↑)		USR (%, ↓)		SER (%, ↑)		Rev (%, ↑)	
	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out
EHRWorld-4B	48.37	53.91	66.44	75.58	31.64	42.73	0.086	0.269	51.56	**42.14**	19.11	14.33	25.87	45.95
EHRWorld-4B (align)	48.29	54.14	67.85	74.04	32.09	42.49	0.096	0.240	53.31	43.56	18.88	16.86	27.22	44.66
MEDROAM-4B (Ours)	**60.85**	**61.02**	**78.80**	**79.89**	**48.11**	**51.10**	**0.165**	**0.281**	**44.91**	49.97	**31.46**	**20.59**	**52.99**	**56.85**
EHRWorld-8B	46.07	53.15	65.93	74.20	29.71	41.54	0.056	0.274	53.18	40.02	16.18	17.50	29.77	47.02
EHRWorld-8B (align)	47.14	53.33	66.11	72.67	30.54	41.25	0.091	0.279	51.79	**38.62**	18.20	18.99	30.92	46.62
MEDROAM-8B (Ours)	**62.47**	**65.08**	**76.21**	**76.47**	**49.68**	**52.22**	**0.185**	**0.361**	**42.51**	50.24	**31.80**	**20.70**	**57.39**	**62.66**
EHRWorld-14B	45.50	56.20	65.69	76.87	28.69	45.89	0.273	0.525	54.01	**39.13**	22.67	18.68	35.40	53.89
EHRWorld-14B (align)	46.61	56.14	65.69	79.53	29.55	46.54	0.223	0.496	49.92	41.15	21.89	18.54	35.20	53.30
MEDROAM-14B (Ours)	**59.02**	**58.26**	**78.04**	**82.14**	**48.01**	**50.74**	**0.370**	**0.597**	**41.33**	39.97	**30.51**	**19.62**	**58.75**	**65.44**

Obs. 2. MEDROAM narrows this gap with consistent rollout gains across model scales. Compared with EHRWorld (align), MEDROAM improves autoregressive rollout performance across all model sizes in both ID and OOD settings. Specifically, on the ID test set, it improves autoregressive *Numer. S@25* and *Stat F1* against EHRWorld (align) by an average of 1.68 and 3.54 points across model scales, respectively. Moreover, these gains extend to the OOD test set, with an average improvement of 4.49 points in *Stat F1* (Full). Notably, our method does not consistently improve teacher-forced performance and yields slight declines on some metrics. This is consistent with our focus on improving autoregressive rollout robustness and narrowing the gap between teacher-forced and autoregressive performance, rather than explicitly enhancing teacher-forced accuracy. Nevertheless, our MEDROAM largely preserves teacher-forced performance while improving autoregressive predictions, demonstrating its effectiveness in strengthening long-horizon rollout robustness.

4.3 INTERVENTION SENSITIVITY

To assess the intervention sensitivity, we evaluate MEDROAM and other medical LWM baselines on four fine-grained probing tasks (L1–L4), covering 14 held-in and 7 held-out probing pairs. Table 2 summarizes the comparative results, from which we can make the following key observations.

Obs. 3. MEDROAM improves both intervention sensitivity and pharmacological consistency. L1–L3 held-in probing results show that continued trajectory training on the selected critical trajectories yields limited improvements. Compared with EHRWorld, EHRWorld (align) achieves modest gains in the composite score of L1 and L2 tasks, while L3 dose sensitivity improves at 4B and 8B but declines at 14B. In contrast, MEDROAM achieves the highest *Sens*, *Mono*, and *Score* across all three sizes. At 4B, it increases *Sens* from 48.29 to 60.85 and *Mono* from 67.85 to 78.80 over EHRWorld (align), raising the macro-averaged *Score* from 32.09 to 48.11. These results indicate that greater responsiveness is accompanied by stronger agreement with the expected pharmacological direction. Moreover, MEDROAM achieves the highest *Dose-*ρ at every size, increasing the average from 0.137 to 0.240 and indicating more consistent dose–response relationships. Overall, these findings support IRA's effectiveness in aligning numerical responses with pharmacological expectations.

Obs. 4. MEDROAM improves held-in L4 probing scores. Although L1–L3 assess sensitivity to removal and dose changes, they do not distinguish intervention-specific responses from general context sensitivity. Thus, L4 evaluates unrelated, similar, and opposing interventions. Compared with EHRWorld (align), MEDROAM achieves lower *USR* at all three sizes, reflecting a larger proportional increase in positive-probe sensitivity than in control sensitivity. Notably, lower USR does not imply reduced responsiveness to negative controls, as absolute control sensitivity also increases (Table 18). For similar substitutions, MEDROAM achieves the highest *SER* across all sizes. Among eligible instances with changed predictions under removal, a larger proportion show nonzero substitution responses smaller than removal responses. This does not imply smaller absolute substitution effects (Figure 12). For opposing substitutions, MEDROAM achieves the highest *Rev*, exceeding EHRWorld (align) by 23.55–26.47 points. Since IRA does not explicitly supervise dose variation or opposing substitutions, these gains suggest benefits beyond the intervention contrasts used in training.

Obs. 5. MEDROAM also improves performance on held-out intervention-response pairs. The preceding analyses evaluate the 14 held-in pairs on different patients. To examine generalization

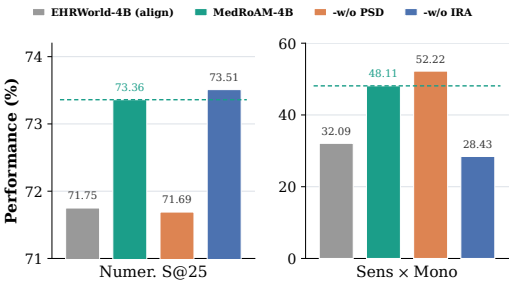


Figure 3: **Ablation of PSD and IRA.**

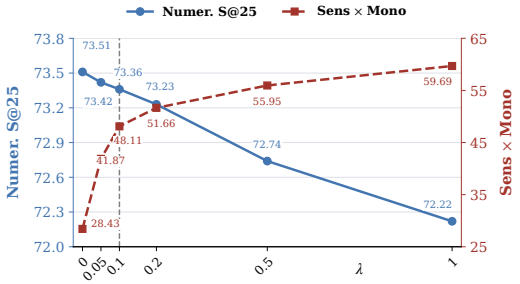


Figure 4: **Effect of alignment strength λ.**

beyond them, we further evaluate pairs excluded from IRA supervision. As shown by the "H-Out" results in Table 2, MEDROAM improves most L1–L4 metrics across model scales. Specifically, it improves the L1–L2 composite score over EHRWorld (align) by 8.61, 10.97, and 4.20 points at 4B, 8B, and 14B, respectively. However, *USR* increases at 4B and 8B and decreases only slightly at 14B, indicating that improved responsiveness to positive probes relative to controls does not consistently extend to held-out relations. This may reflect IRA's supervision of similar substitutions without explicit constraints on unrelated interventions. In future work, we will explore incorporating unrelated and opposing interventions into the supervision. Overall, these results support the generalization beyond the supervised pairs, while highlighting remaining limitations in held-out specificity.

4.4 ABLATION STUDY

We examine the contributions of PSD and IRA using the 4B model: "-w/o PSD" replaces $\mathcal{L}_{\text{PSD}}$ with standard language modeling, yielding $\mathcal{L}_{\text{LWM}} + \lambda\mathcal{L}_{\text{IRA}}$, while "-w/o IRA" uses PSD alone. For a controlled comparison, we match the expected IRA supervision frequency in "-w/o PSD" to that of the full model. We evaluate ID autoregressive accuracy using *Numer. S@25* and held-in intervention sensitivity using *Sens*×*Mono*, with EHRWorld-4B (align) as a reference. As shown in Figure 3, each component primarily improves its targeted capability, *i.e.*, removing PSD reduces rollout accuracy, whereas removing IRA weakens intervention sensitivity. Moreover, "-w/o PSD" achieves a higher intervention sensitivity score than the full model (52.22 vs. 48.11), but lower rollout accuracy (71.69 vs. 73.36). These results reveal a trade-off between the two capabilities. The full model balances these objectives and outperforms EHRWorld-4B (align) on both metrics. We further examine this trade-off by varying the alignment weight $\lambda \in \{0, 0.05, 0.1, 0.2, 0.5, 1\}$ using the same metrics. As shown in Figure 4, increasing λ improves intervention sensitivity but reduces rollout accuracy, indicating that stronger pharmacological alignment can come at the expense of trajectory prediction. Our default setting of $\lambda = 0.1$ offers a favorable trade-off between rollout accuracy and intervention sensitivity.

☞ **Note.** Due to space limitations, we provide related work, data curation, and experimental details in Appendices A–C. Appendix D presents additional experimental evidence through evaluations on other LLM backbones, multiple refinement seeds, and patient-level paired bootstrap confidence intervals. It also compares MEDROAM with scheduled sampling and knowledge prompting, and examines individual components, hyperparameter sensitivity, training data size, and rollout horizons. Further discussion investigates agreement between the IRA surrogate and decoded responses, and clarifies the limitations of our work. Please refer to the Appendix for more details.

5 CONCLUSION

In this paper, we identify and address two limitations of medical LWMs: limited rollout robustness and weak intervention sensitivity. We propose MEDROAM, which jointly optimizes Privileged Self-Distillation (PSD) and Intervention Response Alignment (IRA). PSD uses supervision from ground-truth histories to improve robustness to accumulated rollout errors while anchoring the model's original behavior. IRA incorporates pharmacological knowledge to encourage appropriate numerical responses to intervention changes. Experiments across Qwen3 models demonstrate improved autoregressive rollout accuracy on both ID and OOD data while largely preserving teacher-forced performance. Evaluations using 14 held-in and 7 held-out pairs across 4 probing tasks further show improved intervention sensitivity and greater consistency with pharmacological expectations.

AI USE STATEMENT

We used proprietary large language models (LLMs) as writing assistants to polish language and improve clarity. We also used LLMs for data curation, specifically to identify intervention-response pairs and assist in constructing probing tasks, as described in Appendix B. The resulting outputs were reviewed and filtered using the verification procedures described in that appendix. The research ideas, experimental design, and analyses were developed by the authors. We did not use LLMs to generate research ideas or search for related work. The authors take responsibility for the final content of this work, including AI-assisted text and data.

ETHICS AND REPRODUCIBILITY STATEMENTS

Ethics. We take ethical considerations very seriously and strictly adhere to the ICLR Ethics Policy. This work studies medical language world modeling through retrospective evaluations of rollout accuracy and intervention sensitivity. We use de-identified records from MIMIC-IV and the eICU-CRD Demo dataset in accordance with their respective access requirements and data-use terms. Our evaluations assess predictive performance and pharmacological response consistency, and they do not establish patient-specific causal effects or clinical benefit. Models trained on observational EHRs may inherit biases in patient populations and treatment practices, and improved probe performance does not establish suitability for clinical deployment.

Reproducibility. We describe our method in Section 3 and provide training hyperparameters, inference settings, and evaluation metrics in Appendix C.3. Moreover, Appendices B.4 and B.5 detail the construction of intervention–response pairs and L1–L4 probing tasks, respectively. To facilitate reproducibility, ***we provide our training code and data processing scripts in the supplementary material***. Regarding data availability, MIMIC-IV requires authorized access through PhysioNet, while the eICU-CRD Demo dataset is publicly available under its data-use terms. Note that we do not redistribute patient-level data.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Harry Amad and Mihaela van der Schaar. DT^2: Decision-targeted digital twins. In *Forty-third International Conference on Machine Learning*, 2026.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela Van Der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pp. 301–318. PMLR, 2016a.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016b.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. Clarity: Medical world model for guiding treatment decisions by simulating context-aware disease trajectories. In *European Conference on Computer Vision*. Springer, 2026.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Pujun Feng, Xiaoyu Guo, Seyed Ehsan Saffari, Min Hun Lee, Siew-Kei Lam, Erik Cambria, Xibin Sun, Yangtao Zhou, Tong Yang, Xiaoyu Zhang, et al. From static risk to dynamic trajectories: Toward world-model-inspired clinical prediction. *arXiv preprint arXiv:2605.16927*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 81785–81830, 2025.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International conference on machine learning*, pp. 15293–15329. PMLR, 2022.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Linjie Mu, Zhongzhen Huang, Yannian Gu, Shengqian Qin, Shaoting Zhang, and Xiaofan Zhang. Ehrworld: A patient-centric medical world model for long-horizon clinical trajectories. *arXiv preprint arXiv:2602.03569*, 2026.

Tom J Pollard, Alistair EW 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(1):180178, 2018.

Mohammad Areeb Qazi, Maryam Nadeem, and Mohammad Yaqub. Beyond generative ai: World models for clinical prediction, counterfactuals, and planning. *arXiv preprint arXiv:2511.16333*, 2025.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Numan Saeed, Salma Hassan, Shadab Khan, Mohammad Areeb Qazi, Klaus H Maier-Hein, Salman Khan, and Mohammad Yaqub. Medical world model: From passive prediction to active simulation in medicine. 2026.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Siri Willems, James Butterworth, Lore Goetschalckx, Peter Vrancx, Philippe Modard, Elke Giets, and Ludovic Denoyer. Future querying: Can llms serve as implicit medical world models? *arXiv preprint arXiv:2608.23248*, 2026.

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

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, 2025.

Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, et al. Pan: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057*, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Qianyi Xu, Gousia Habib, Feng Wu, Dilruk Perera, and Mengling Feng. Meddreamer: Model-based reinforcement learning with latent imagination on complex ehrs for clinical decision support. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1693–1704, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, et al. Medical world model: Generative simulation of tumor evolution for treatment planning. *arXiv preprint arXiv:2506.02327*, 2025b.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026a.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, et al. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026b.

APPENDIX

A RELATED WORK

A.1 EHR MODELING AND MEDICAL WORLD MODELS

Electronic health records (EHRs) document patient histories, clinical measurements, and treatments over time, providing a foundation for modeling disease progression and predicting clinical outcomes. Early approaches use recurrent networks (Choi et al., 2016a;b) to encode longitudinal histories, while subsequent BERT-based models, such as BEHRT (Li et al., 2020) and Med-BERT (Rasmy et al., 2021), learn contextualized patient representations through pretraining for disease prediction. Recently, large language models (LLMs) have expanded the use of medical knowledge and textual reasoning in healthcare (Singhal et al., 2023). For example, KARE (Jiang et al., 2025) enriches patient contexts through knowledge graph retrieval and uses LLM reasoning to predict mortality and readmission. While effective, these approaches primarily predict outcomes from recorded histories. They can answer *what will happen to this patient*, but not *what would happen under a different intervention*, which requires modeling how patient states evolve explicitly.

This motivates medical world models (WMs), which learn state transitions conditioned on the patient history and a given action (Ha & Schmidhuber, 2018; Hafner et al., 2019). By repeatedly predicting the next state, a WM can simulate trajectories under different intervention sequences (LeCun, 2022). Existing medical WMs mainly differ in their state representations. For instance, MeWM (Yang et al., 2025b) generates images of tumor evolution under treatment conditions, while CLARITY (Ding et al., 2026) predicts disease trajectories in a latent space that incorporates temporal and clinical context. More recently, medical WMs built on LLM backbones represent clinical states and interventions as text, offering the potential to draw on extensive pretrained medical knowledge to improve simulation accuracy. EHRWorld (Mu et al., 2026) follows this approach, organizing longitudinal EHRs into sequences of interventions and observations to train medical language world models (LWMs). Future Querying (Willems et al., 2026) explores a related direction through time-indexed questions about future patient states using clinical documentation. Following EHRWorld's trajectory formulation, our work focuses on assessing and improving two key capabilities of medical LWMs, *i.e.*, sustaining accurate predictions over long rollouts and responding appropriately to intervention changes.

A.2 MITIGATING EXPOSURE BIAS IN AUTOREGRESSIVE MODELS

Autoregressive language models are typically trained on ground-truth prefixes but condition on their own predictions at inference. This *training-inference mismatch*, commonly termed "exposure bias", causes early errors to propagate through subsequent predictions (Ranzato et al., 2015; Bengio et al., 2015). In medical LWMs, feeding predicted physiological values back into the history makes this mismatch particularly relevant to long-horizon rollout accuracy (Mu et al., 2026). To mitigate exposure bias, existing approaches expose models to their own predictions during training. For instance, scheduled sampling (Bengio et al., 2015) replaces selected ground-truth inputs with self-generated tokens. On-policy distillation (Agarwal et al., 2024) provides teacher supervision on on-policy student-generated sequences. Recent on-policy self-distillation methods obtain teacher guidance by enriching the model's conditioning context. OPSD (Zhao et al., 2026a) uses reference solutions as privileged information to supervise the model's own reasoning trajectories, while SDFT (Shenfeld et al., 2026) uses demonstration-conditioned self-distillation to learn new capabilities while reducing forgetting. SDPO (Hübotter et al., 2026) further uses environmental feedback to construct token-level self-distillation signals. These methods demonstrate how additional privileged information can guide learning on the student's on-policy outputs.

Despite their effectiveness, these methods are difficult to directly apply to medical LWMs. Since EHR trajectories often span many steps, on-policy autoregressive sampling is computationally expensive during training. To address this challenge, we propose an off-policy Privileged Self-Distillation (PSD) approach. Specifically, we generate and cache student rollout histories offline, then use them as student inputs during training. A frozen teacher conditioned on recorded histories provides privileged supervision, guiding the student toward accurate simulation from imperfect histories. Moreover, to mitigate forgetting, we mix recorded and cached rollout histories as student inputs, explicitly

anchoring predictions on recorded histories to the frozen teacher. By doing so, our approach can efficiently improve rollout robustness while preserving teacher-forced performance.

A.3 Intervention Sensitivity in World Models

Accurate trajectory prediction alone does not ensure appropriate responses to intervention changes, as a medical WM may rely primarily on patient history and make little use of intervention inputs. Existing work uses various training objectives to model treatment-dependent outcomes and enable intervention-aware simulation. For instance, Counterfactual Recurrent Networks (Bica et al., 2020) and the Causal Transformer (Melnychuk et al., 2022) learn balanced representations to address observed time-varying confounding when estimating counterfactual outcomes, while DT^2 (Amad & van der Schaar, 2026) trains world models to preserve policy-value rankings obtained through offline policy evaluation. These methods target causal outcome estimation or policy comparison and are primarily designed for task-specific models rather than pretrained LLM backbones.

However, medical LWMs introduce an additional challenge. Their strong pretrained language priors may encourage plausible predictions without sufficient attention to the intervention specified in context. To address this issue, we propose Intervention Response Alignment (IRA), which uses curated intervention-response pairs with pharmacologically supported effect directions. Specifically, IRA constrains differentiable numerical surrogates derived from value-token probabilities to align predicted responses with pharmacological expectations. This encourages the model to attend to intervention inputs when predicting patient states. Notably, IRA does not aim to encode all pharmacological knowledge, but to provide a training signal that strengthens intervention sensitivity. Evaluations on seven held-out intervention-response pairs show that these benefits extend beyond directly supervised relations, improving intervention responsiveness and consistency with pharmacological expectations.

B Data Processing Details

B.1 From MIMIC-IV to the Longitudinal ICU Trajectory Corpus

The event sequence of each trajectory is constructed through a deterministic pipeline over MIMIC-IV tables, allowing every observation and intervention to be traced to its source table. In general, we follow the data curation procedure of EHRWorld (Mu et al., 2026). However, we focus primarily on the ICU module rather than the hospital module and do not extract patient diagnoses from discharge summaries, avoiding the use of future information from these summaries for state prediction. Specifically, the cohort includes each patient's *first* ICU stay lasting at least 24 hours and retains only the first 72 hours after admission, yielding 54,551 stays. All content comes from structured tables, without using clinical notes or language models during dataset construction.

From the ICU module, we use `icustays` for cohort definition and care units, `chartevents` for vital signs, `inputevents` for drug infusions and boluses, `procedureevents` for procedures, and `outputevents` for urine output. Vital signs include mean arterial pressure, heart rate, respiratory rate, SpO_2, and temperature, selected by `itemid`. Procedure names are obtained from `d_items`. From the hospital module, we use `labevents` and select twenty `itemids` mapping to nineteen analytes (serum and blood-gas glucose are merged into one channel), covering electrolytes, renal markers, glucose, blood counts, coagulation, and lactate. Each laboratory value is rendered with its recorded reference interval and marked `abnormal` if it falls outside that interval. We select analytes by `itemid` rather than name, as a single analyte may have multiple identifiers in MIMIC-IV, and name-based matching could yield a different set of measurements. The clinical context prefixed to each trajectory contains three fields obtained from structured tables. Age is computed from `patients.anchor_age`, adjusted by the difference between the ICU admission year and `anchor_year`. Gender is obtained from the same table, and the admitting ICU type is taken from `icustays.first_careunit`. Finally, each trajectory is rendered as plain text, with events ordered chronologically by their recorded `charttime`. A `<patient>` header contains the clinical context, followed by event blocks delimited by `<step charttime>`. Within each block, observations are formatted as `[inquiry] Name -> value unit [ref lo-hi] flag`, and interventions as `[intervention] drug rate unit`. Table 3 provides an example.

Table 3: **An example from a real ICU trajectory** illustrating the state–action–state sequence.

```
<patient>
Age: 80, Gender: M, ICU: Cardiac Vascular Intensive Care Unit (CVICU)

<step 2138-02-04T11:23:00>
[inquiry] Glucose -> 297 mg/dL [ref 70.0-105.0] abnormal
[inquiry] Lactate -> 2.2 mmol/L [ref 0.5-2.0] abnormal

<step 2138-02-04T12:44:00>
[inquiry] Glucose -> 310 mg/dL [ref 70.0-105.0] abnormal
[inquiry] Lactate -> 4.1 mmol/L [ref 0.5-2.0] abnormal
                          ... (3 steps omitted) ...

<step 2138-02-04T14:33:00>
[inquiry] MAP -> 70 mmHg

<step 2138-02-04T14:33:00>            <-- critical decision step
[intervention] Insulin - Regular 3 units/hour
[intervention] Epinephrine 0.05 mcg/kg/min
[intervention] Propofol 40.064 mcg/kg/min
[intervention] IV fluids 52 mL/hr
                       ... (6 more interventions) ...

<step 2138-02-04T14:34:00>
[inquiry] Heart Rate -> 80.0 bpm
[inquiry] Respiratory Rate -> 16.0 insp/min
[inquiry] SpO2 -> 100.0 %
                          ... (7 steps omitted) ...

<step 2138-02-04T15:19:00>
[inquiry] Glucose -> 224 mg/dL [ref 70.0-105.0] abnormal

<step 2138-02-04T15:21:00>
[intervention] LR 30 mL/hour
[intervention] Insulin - Regular 6.998 units/hour
[intervention] Insulin - Regular 4 units
[intervention] IV fluids 7 mL/hr
                        ... (trajectory continues) ...
```

Two design choices preserve the recorded temporal ordering of observations and interventions. First, we order events by their recorded `charttime` rather than grouping them into hourly bins. Hourly binning obscures within-hour ordering and, depending on the serialization rule, may place an observation before an intervention that preceded it in the original records. Event-level ordering reduces such temporal inconsistencies, although recorded timestamps do not establish causal relationships or precisely identify when physiological effects occur.

Second, for events sharing a timestamp, we place `[inquiry]` before `[intervention]` as a conservative serialization convention. This avoids treating simultaneous observations as responses to the intervention without evidence of their temporal order. Under this convention, same-time inquiries belong to the pre-action history h_t, while s_t contains target observations recorded after a_t. Within each timestamp, observations follow a fixed clinical order to ensure consistent serialization across the corpus. To reduce redundancy in frequently recorded vital signs, we retain a reading only when it differs from the last retained value by more than a channel-specific threshold or when more than one hour has elapsed. This reduces sampling density during stable periods while retaining changes that exceed the threshold. All other channels remain uncompressed. For medication events, continuous infusions are rendered as rate lines and single administrations as dose lines, while carrier fluids are grouped into a single `IV fluids` entry. Lastly, the constructed corpus is split at the patient level, with 5% held out under a fixed seed, resulting in 51,624 training, 200 development, and 2,727 test stays. More specifically, sequence lengths have a median of about 4.4k tokens and a 90th percentile of about 7.5k, compatible with the 16k training cutoff.

B.2 Cross-Center OOD Test Set

Training and testing on MIMIC-IV alone can only show that a model works under the recording conventions of a single hospital. To test whether it has learned transferable physiological regularities, we additionally construct a cross-center OOD test set from eICU-CRD (Demo v2.0.1)[1]. Specifically, eICU is collected across multiple hospitals in the United States, and differs from MIMIC-IV in its acquisition systems, the granularity of medication records, and its conventions for laboratory naming and reference ranges. Thus, it is a genuine cross-center transfer test. The key constraint is that this test set is serialized into a text format *identical field by field* to that of the MIMIC-IV corpus, so that the same world model can be evaluated directly without retraining or adaptation.

The cohort definition is aligned strictly with our training set. We take each patient's first ICU stay by `uniquepid`, require a length of stay of at least 24 hours, and retain only the first 72 hours, yielding **1,333** stays with an average of 445 observation steps each (median 436, 90th percentile 709). In particular, mapping eICU into our format requires handling four systematic differences:

Time. For de-identification, eICU provides only minute offsets relative to ICU admission rather than absolute timestamps. We synthesize timestamps by adding these offsets to a fixed reference time, so that the `<step>` headers take a consistent form. The pseudo-timestamps preserve the ordering and time intervals between steps, which are the temporal information relevant to trajectory modeling.

Observation channels. Vital signs are taken from `vitalPeriodic` and `vitalAperiodic` tables, with the non-invasive mean arterial pressure mapped to MAP, and the same compression thresholds as MIMIC are applied. Urine output is derived from `intakeOutput` entries and converted to mL/hr from consecutive time differences. Laboratory items are normalized to the same channels the world model predicts through an explicit name mapping (for example `BUN` to `Urea Nitrogen` and `WBC x 1000` to `White Blood Cells`).

Reference ranges. The eICU `lab` table provides neither reference ranges nor abnormality flags, whereas `[ref lo-hi]` and `abnormal` are constituent parts of an inquiry line in our format. We attach the modal reference range estimated from the MIMIC training split for each channel and compute the abnormality flag accordingly, and we also unify unit strings to the MIMIC convention (*e.g.*, mEq/L rather than mmol/L for electrolytes, such as sodium and potassium, which are numerically equivalent). The intent is to confine the distribution shift to physiology and prescribing rather than to surface formatting, since otherwise the evaluation would measure how the tokenizer reacts to unfamiliar unit strings rather than cross-center generalization.

Interventions. We use only continuous infusions from the `infusiondrug` table, whose `drugname` field encodes the unit in parentheses, as in `Norepinephrine (mcg/min)`. We split this into a drug name and a unit before rendering it as a rate line. The `medication` and `treatment` tables are deliberately excluded, since the former records orders rather than administrations and the latter is a hierarchical taxonomy of treatment categories. Neither aligns with the event-level administration records of MIMIC, and forcing them in would introduce formatting noise rather than genuine distribution shift.

We acknowledge that two limitations of this test set remain. First, because only infusions are retained, the action space of eICU trajectories is narrower than that of MIMIC, with procedures, blood products, and bolus administrations absent. Hence, this set primarily tests the cross-center rollout accuracy and is not suited for evaluating intervention sensitivity. Second, the reference ranges are borrowed from MIMIC rather than self-reported by each eICU center. This is a deliberate trade-off that reduces surface-level shift, but it also means that the abnormality flags reflect MIMIC's criteria.

B.3 Critical Intervention Steps

Most intervention steps in an ICU trajectory involve routine fluids, nutrition, or maintenance medications. To focus refinement training on clinically meaningful treatment changes, we define a critical intervention step by the following two conditions.

Axis A: initiation of a high-risk treatment class. The step must introduce at least one high-risk treatment class not seen earlier in the episode. The classes cover vasopressors, diuretics, insulin infusions,

[1] https://physionet.org/content/eicu-crd-demo/2.0.1/

antiarrhythmics, antibiotics, blood products, airway and ventilation procedures, and renal replacement therapy. The antibiotic list is read directly from entries with `category=Antibiotics` in the `d_items` table rather than enumerated by hand. We require the *initiation* of a class rather than its mere presence, since the latter is dominated by maintenance dosing.

Axis B: an acute and actionable state. Within a lookback window at the intervention time (6 hours by default), the most recent value of a monitored channel must cross an acute and actionable threshold. Table 4 lists the full thresholds for different channels. Each critical step also records the specific state label that triggered it, so that results can be stratified by trigger type.

Table 4: **Acute thresholds for identifying Axis B critical steps**. All 25 channels remain prediction targets, and only those listed contribute to critical-step identification. Units follow MIMIC-IV.

Channel	Unit	Threshold	Label
MAP	mmHg	< 65	Hypotension
Urine	mL/hr	< 30	Oliguria
SpO_2	%	< 90	Hypoxia
Heart rate	bpm	> 130	Tachycardia
		< 45	Bradycardia
Respiratory rate	insp/min	> 28	Tachypnea
Temperature	°C	> 38.5	Fever
		< 35.5	Hypothermia
Glucose	mg/dL	> 250	Hyperglycemia
		< 60	Hypoglycemia
Potassium	mEq/L	> 6.0	Hyperkalemia
		< 3.0	Hypokalemia
Sodium	mEq/L	> 155	Hypernatremia
		< 125	Hyponatremia
Lactate	mmol/L	> 4.0	Hyperlactatemia
Bicarbonate	mEq/L	< 15	Acidosis
Hemoglobin	g/dL	< 7.0	Severe anemia

As a result, on the test set, only 2,504 intervention steps satisfy both conditions. Applying the same criterion to the training split yields 24,216 episodes containing at least one critical step. Notably, in the training split, matching critical interventions to the 14 curated pairs (mentioned in the following content) and requiring a parseable target response within the corresponding window yields 8,609 trajectories containing 12,801 intervention-response instances. Specifically, the "furosemide–urine" pair accounts for 6,821 instances, more than half of the total. To reduce this imbalance, we subsample this pair to 1,463 instances, approximately matching the next most frequent pair, "norepinephrine–MAP", with 1,472 retained instances. Trajectories with no remaining instances are discarded. After offline rollout generation, we also exclude failed generations, mismatched value-span counts, and inconsistent abnormality markers to preserve the validity of paired histories. The resulting refinement subset contains 5,646 trajectories, which form the training corpus for MEDROAM. Notably, these criteria apply to refinement-data selection. For probing evaluation, we use the target drug's first administration in each eligible test trajectory, regardless of whether it occurs at a critical step.

B.4 CURATING INTERVENTION–RESPONSE PROBING PAIRS

Evaluating intervention responses requires drug–channel pairs with externally supported effect directions. In observational ICU data, prescribing is confounded by indication, so a treatment may be associated with the abnormal measurement that prompted its administration. Therefore, we establish probe directions using external pharmacological evidence through four screening stages, including candidate generation, model agreement and mechanism filtering, drug-label adjudication, and training-data support. A subsequent drift-adjusted diagnostic characterizes the retained pairs without affecting their inclusion. Notably, these probes assess pharmacological consistency under specified conditions rather than patient-specific counterfactual accuracy. In this section, we first describe how we curate the 14 intervention–response pairs used for IRA training and held-in probing, then introduce the seven additional pairs used for held-out evaluation.

B.4.1 INTERVENTION–RESPONSE PAIRS FOR TRAINING AND HELD-IN EVALUATION

Stage 1. Candidate generation. We identify 56 drug names at critical steps in the training split, corresponding to 53 distinct active ingredients. Counts through Stage 3 refer to drug-name–channel pairs because labels are retrieved separately for each name. Naming variants are merged by active ingredient after label adjudication. For each name, we ask three LLMs (`Qwen3.8-27B`, `MedGemma-27B`, and `DeepSeek-V4-Flash`) to identify which of the 25 monitored channels the drug directly affects, along with the expected direction, time course, and mechanism. The prompt provides no candidate answers and explicitly allows an empty response to discourage unsupported proposals. This defines a search space of $56 \times 25 = 1{,}400$ drug–channel pairs.

Table 5: **The 14 held-in label-supported pharmacological probes**, grouped by drift-adjusted training responses. *Dir* denotes the expected direction, and $[\omega_{\text{lo}}, \omega_{\text{hi}}]$ gives the query window in hours. *Src* indicates label-reported timing (LBL), label information with an inferred or human-selected horizon (LBL*), cross-channel timing transfer (INF), or drift-adjusted lag curves (DAT). |ES| is the largest label-consistent adjusted deviation, normalized by the channel's median absolute inter-measurement change; "—" denotes no such peak. Groups are descriptive, not inclusion criteria. n_{tr} counts episodes with the target drug at an eligible critical step across the full training split before refinement filtering and balancing, once per probe. n_{te} counts first-administration test instances, one per eligible episode, without requiring a critical step. Links point to DailyMed labels.

Drug	Channel	Dir.	$[\omega_{\text{lo}}, \omega_{\text{hi}}]$	Src.	|ES|	n_{tr}	n_{te}	Label
Larger adjusted response								
furosemide	Urine	↑	[0.08, 2.0]	LBL	9.35	7,113	805	[DailyMed]
amiodarone	Heart Rate	↓	[1.0, 24.0]	DAT	3.42	1,140	128	[DailyMed]
dopamine	Heart Rate	↑	[0.08, 0.17]	LBL	1.27	454	54	[DailyMed]
furosemide	Chloride	↓	[1.0, 24.0]	DAT	1.16	7,113	805	[DailyMed]
diltiazem	MAP	↓	[0.08, 3.0]	LBL*	0.58	755	74	[DailyMed]
insulin	Glucose	↓	[0.35, 2.0]	LBL*	0.35	6,056	1,093	[DailyMed]
Weak adjusted response								
phenylephrine	MAP	↑	[0.08, 0.33]	LBL*	0.10	4,044	573	[DailyMed]
furosemide	Sodium	↓	[12.0, 24.0]	DAT	0.06	7,113	805	[DailyMed]
furosemide	Potassium	↓	[8.0, 12.0]	DAT	0.05	7,113	805	[DailyMed]
furosemide	Magnesium	↓	[4.0, 8.0]	DAT	0.01	7,113	805	[DailyMed]
No label-consistent peak								
norepinephrine	MAP	↑	[0.08, 0.50]	LBL*	—	3,505	442	[DailyMed]
epinephrine	MAP	↑	[0.08, 0.33]	LBL*	—	697	128	[DailyMed]
epinephrine	Heart Rate	↑	[0.08, 0.33]	INF	—	697	128	[DailyMed]
vasopressin	MAP	↑	[0.08, 0.33]	LBL*	—	407	123	[DailyMed]

Stage 2. Agreement and mechanism filtering. We retain a pair only if all models agree on its direction and each provides a specific mechanism. Responses that merely state that a drug affects or modulates a channel without explaining the mechanism are excluded. These criteria retain **33** pairs.

Stage 3. Adjudication against drug labels. We retrieve structured product labels through the FDA's openFDA service[2] and examine the adverse reactions, clinical pharmacology, and warnings and precautions sections. We exclude the indications section to avoid interpreting the treated condition as a drug effect and prioritize intravenous formulations to match the evaluated administration route. Evidence retrieval and directional adjudication are performed separately. Specifically, we retrieve passages using channel-specific keywords that do not encode an expected direction, then assess their meaning in context. This distinguishes physiological effects from conditional statements, interactions, pharmacokinetics, and assay interference. Each passage set is independently assessed by `Qwen3.8-27B` at temperature 0 and a human adjudicator. Each assessment returns one of four verdicts, *supports*, *contradicts*, *conditional*, or *not_stated*, together with a verbatim evidence passage when available and a flag indicating formulation and route compatibility. We retain a pair only when both assessments support the proposed direction and confirm compatibility. Effects requiring a particular comorbidity or co-administered drug are classified as conditional and excluded. This stage retains **18** label-supported pairs.

Notably, contextual adjudication can reject directions proposed during candidate generation. Both adjudicators returned *contradicts* for "lidocaine–heart rate" after reviewing passages describing minimal hemodynamic effects and toxicity-related bradycardia, and the formulation compatibility flag was negative. Both also rejected the proposed increase for "norepinephrine–heart rate" because the label describes reflex vagal slowing and lists bradycardia among common adverse reactions.

Stage 4. Training-data support. We randomly sample 15,000 training episodes and retain candidates whose target drug is administered in at least 275 episodes, regardless of whether administration occurs at a critical step. Each episode is counted once per drug. This prevalence of 1.83% corresponds to approximately 50 expected episodes in the 2,727-episode test set under comparable prevalence, but

[2] https://open.fda.gov

Table 6: **The seven held-out intervention-response probes.** These pairs are excluded from IRA supervision. *Dir.* denotes the expected effect direction, $[\omega_{\text{lo}}, \omega_{\text{hi}}]$ specifies the query window in hours, and n_{te} counts test probing instances. Links refer to product labels on DailyMed.

Drug	Channel	Dir.	$[\omega_{\text{lo}}, \omega_{\text{hi}}]$	n_{te}	Label
nitroglycerin	Heart Rate	↑	[1.0, 12.0]	361	[DailyMed]
fentanyl	Heart Rate	↓	[1.0, 24.0]	813	[DailyMed]
midazolam	MAP	↓	[1.0, 24.0]	323	[DailyMed]
magnesium sulfate	MAP	↓	[1.0, 4.0]	1,147	[DailyMed]
nicardipine	Heart Rate	↑	[1.0, 2.0]	168	[DailyMed]
dexmedetomidine	MAP	↓	[0.5, 8.0]	197	[DailyMed]
hydralazine	Heart Rate	↑	[0.5, 8.0]	263	[DailyMed]
Total				3,272	

does not guarantee a minimum evaluation count. This stage retains the **14** pairs listed in Table 5. Separately, the n_{tr} values in that table summarize critical-step support in the full training split before refinement filtering and balancing, rather than the counts used for the 275-episode threshold. An episode may contribute to multiple probes, but is counted at most once per probe. The refinement subset is subsequently constructed through response matching, downsampling, and rollout validity checks. Thus, n_{tr} does not represent the final refinement sample size, and exposure frequency indicates data support rather than causal identifiability.

B.4.2 Intervention–Response Pairs for Held-Out Evaluation

The 14 pairs selected above are used for both IRA training and subsequent intervention sensitivity evaluation. To examine generalization beyond these supervised relations, we construct seven additional intervention-response pairs excluded from IRA supervision. In general, we follow the same verification principles, using agreement across models and mechanism filtering, followed by FDA-label adjudication to verify the proposed pharmacological directions. We then assess training-data support to ensure sufficient supporting episodes.

The main difference lies in candidate generation. Instead of restricting candidates to interventions appearing at critical steps, we consider the intervention set across the full training split. We exclude routine and simple interventions, and deliberately focus on challenging relations involving secondary effects, baroreflex responses, or weak effects, while also considering frequently measured channels. All selected pairs are distinct from the 14 held-in pairs and absent from IRA's directional supervision set. Thus, this evaluation examines generalization to intervention-response relations not directly supervised during refinement. The resulting seven pairs are summarized in Table 6.

B.5 L1–L4 Probing Tasks

Using the curated intervention-response pairs, we construct four levels of probing tasks to assess whether predictions respond to intervention changes in pharmacologically appropriate ways. For both held-in and held-out evaluation, we search the full test set for episodes containing the probe drug and construct one instance per probe-episode pair at the drug's first administration. The tasks share the query construction and decoding protocol described below.

L1. Removal responsiveness. We first test whether the model uses the intervention when predicting its associated response. For each instance, we compare the decoded numerical predictions under the original intervention and its removal, denoted by $v^{\text{o}}_{\text{dec}}$ and $v^{\text{r}}_{\text{dec}}$. We delete only the order lines corresponding to the target drug at its first administration step, preserving other interventions at that step and the shared pre-administration history. The metric *Sens* is the percentage of instances with $v^{\text{o}}_{\text{dec}} \neq v^{\text{r}}_{\text{dec}}$, measuring how often removal changes the prediction.

L2. Directional consistency. A changed prediction need not reflect the expected drug effect. Thus, using the same instances as L1, we further test whether removal reverses the label-supported direction. Let $\sigma_p \in \{-1, +1\}$ denote the expected direction of probe p under the given intervention. Among instances with changed predictions, Mono_p is the percentage satisfying $\sigma_p(v^{\text{o}}_{\text{dec}} - v^{\text{r}}_{\text{dec}}) > 0$. We compute $\text{Score}_p = \text{Sens}_p \times \text{Mono}_p/100$ for each probe to jointly assess responsiveness and

directional consistency, then report $\text{Score} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \text{Score}_p$ over the probe set $\mathcal{P}$. That is, Score is the macro-average of per-probe scores, rather than the product of macro-averaged Sens and Mono.

L3. Dose sensitivity. To examine whether the model responds to treatment intensity, we multiply the recorded dose by $f \in \{0.25, 0.5, 1, 2, 4\}$ while keeping the drug identity and remaining context fixed. For each instance, we compute the Spearman rank correlation between dose factors and predicted values. We multiply this correlation by σ_p so that positive values indicate agreement with the expected pharmacological direction. The metric *Dose-ρ* is the median adjusted correlation within each probe. Notably, ρ denotes rank correlation, distinct from the PSD history-mixing probability.

L4. Intervention specificity. We next examine whether sensitivity reflects the pharmacological relationship between an intervention and the target channel. This task includes unrelated interventions, similar substitutes, and opposing substitutes. To evaluate responses to unrelated interventions, we construct negative controls by screening candidate pairs against drug labels using the same direction-independent channel keywords and label sections as in positive-probe adjudication. Positive probes require explicit label support, whereas negative controls require no documented effect on the target channel. Thus, we assess each drug–channel pair individually rather than excluding entire drugs or channels. Notably, the absence of a documented effect does not establish zero physiological effect.

For held-in evaluation, we consider drugs for which all three elicitation models report no effect on any monitored channel. We select the two most frequent at critical training steps, "cefazolin" and "levofloxacin", and pair them with the eight positive-probe channels, yielding 16 candidates. Manual label review retains three pairs whose keyword matches concern only assay interference, renal elimination, or hydration advice, but excludes five with documented effects, namely MAP for both drugs and glucose, heart rate, and potassium for levofloxacin. This leaves 11 held-in controls. For held-out evaluation, the probes cover only heart rate and MAP, leaving one applicable held-in control. Therefore, we screen 20 additional, manually selected pairs using a stricter rule that excludes any pair with a retrieved keyword match. We retain eight controls, comprising six heart-rate pairs and two MAP pairs (Table 7). We apply the L1 removal operation and compute $\text{USR} = \overline{\text{Sens}}_{\text{ctrl}}/\overline{\text{Sens}}_{\text{pos}}$, independently macro-averaging control and positive-probe sensitivities. Held-in evaluation uses 11 controls and 14 positive probes, while held-out evaluation uses 8 controls and 7 positive probes. Note that lower USR indicates less responsiveness to controls relative to target interventions, but not necessarily lower absolute control sensitivity (see Table 18). These comparisons characterize nonspecific responsiveness without establishing a biological null.

Table 7: **Negative controls and query windows for held-in and held-out evaluation.** Notably, windows specify evaluation times, not established zero-effect intervals.

Drug	Channel	$[\omega_{\text{lo}}, \omega_{\text{hi}}]$ (h)
Held-in negative controls		
cefazolin	Chloride	[1.0, 4.0]
cefazolin	Glucose	[0.25, 1.0]
cefazolin	Heart Rate	[0.01, 4.0]
cefazolin	Magnesium	[2.0, 6.0]
cefazolin	Potassium	[1.0, 4.0]
cefazolin	Sodium	[1.0, 4.0]
cefazolin	Urine	[0.5, 2.0]
levofloxacin	Chloride	[1.0, 4.0]
levofloxacin	Magnesium	[2.0, 6.0]
levofloxacin	Sodium	[1.0, 4.0]
levofloxacin	Urine	[0.5, 2.0]
Held-out negative controls		
pantoprazole	Heart Rate	[0.01, 4.0]
pantoprazole	MAP	[0.01, 4.0]
piperacillin	Heart Rate	[0.01, 4.0]
cefepime	Heart Rate	[0.01, 4.0]
acetaminophen	Heart Rate	[0.01, 4.0]
heparin sodium	Heart Rate	[0.01, 4.0]
ceftriaxone	MAP	[0.01, 4.0]
levetiracetam	Heart Rate	[0.01, 4.0]

For similar and opposing substitutions, we use different selection procedures for held-in and held-out probes (Table 8). For held-in probes, we select training-set drugs from the Stage 1 associations (Appendix B.4), requiring unanimous agreement among the three LLMs on the target-channel direction, σ_p for similar substitutes and $-\sigma_p$ for opposing substitutes. Selection prioritizes same-class drugs for similar substitutes, then training-set order frequency. For held-out probes, the original candidate pool has limited coverage, and indirect or reflex-mediated effects require contextual verification. Thus, we manually select substitutes and verify their correctness through openFDA. Each substitute uses its most frequent order string from the training split for held-in probes and the full training set for held-out probes. In both sets, similarity and opposition concern effect direction on the target channel, without assuming equal magnitudes. Let $v^{\text{s}}_{\text{dec}}$ denote the prediction under a similar substitute. For each probe, we retain instances with valid numerical predictions under the original, removal, and similar-substitution arms and with $v^{\text{r}}_{\text{dec}} \neq v^{\text{o}}_{\text{dec}}$. The *similar-drug equivalence*

Table 8: **Substitution arms for L4 probing task.** Substitutes are expected to share or oppose the original probe's direction on the target channel, without assuming equal effect magnitudes. "—" denotes no available substitute. The four "furosemide-electrolyte" probes share the listed substitution.

Probe	Same-direction substitute	Opposite-direction substitute
Held-in substitution arms		
insulin → Glucose	insulin (other formulation)	epinephrine
furosemide → Urine	bumetanide	—
furosemide → electrolytes	bumetanide	—
phenylephrine → MAP	norepinephrine	amiodarone
norepinephrine → MAP	phenylephrine	amiodarone
amiodarone → Heart Rate	esmolol	epinephrine
diltiazem → MAP	amiodarone	norepinephrine
epinephrine → Heart Rate	dopamine	amiodarone
epinephrine → MAP	norepinephrine	amiodarone
dopamine → Heart Rate	epinephrine	amiodarone
vasopressin → MAP	norepinephrine	amiodarone
Held-out substitution arms		
nitroglycerin → Heart Rate	nicardipine hydrochloride	metoprolol tartrate
fentanyl → Heart Rate	dexmedetomidine	epinephrine
midazolam → MAP	propofol	phenylephrine hydrochloride
magnesium sulfate → MAP	nicardipine hydrochloride	phenylephrine hydrochloride
nicardipine → Heart Rate	nitroglycerin	metoprolol tartrate
dexmedetomidine → MAP	propofol	phenylephrine hydrochloride
hydralazine → Heart Rate	nicardipine hydrochloride	metoprolol tartrate

rate (SER) is the percentage of these instances satisfying $0 < |v^{\mathrm{s}}_{\mathrm{dec}} - v^{\mathrm{o}}_{\mathrm{dec}}| < |v^{\mathrm{r}}_{\mathrm{dec}} - v^{\mathrm{o}}_{\mathrm{dec}}|$, macro-averaged across probes. Thus, instances with unchanged predictions under removal are excluded from the denominator. This criterion captures a nonzero substitution response smaller than the removal response, without establishing pharmacological equivalence. Unlike $\mathcal{L}_{\mathrm{equiv}}$, which permits identical surrogate values and uses an adaptive tolerance, SER evaluates decoded values and does not measure training-constraint satisfaction. For an opposing substitute, ***Rev*** is the percentage satisfying $\sigma_p(v^{\mathrm{opp}}_{\mathrm{dec}} - v^{\mathrm{o}}_{\mathrm{dec}}) < 0$, where $v^{\mathrm{opp}}_{\mathrm{dec}}$ denotes the prediction under an opposing substitute, testing an opposite-direction displacement relative to the original prediction. Probes for which no suitable opposing substitute is available are excluded from the calculation of *Rev*.

Shared evaluation settings. Because pharmacological responses develop over different time scales, each probe has a query window $[\omega_{\mathrm{lo}}, \omega_{\mathrm{hi}}]$. As shown in Table 5, two windows use label-reported timing (LBL), six combine label information with an inferred or human-selected horizon (LBL*), one transfers timing from another channel (INF), and five use training-data diagnostics (DAT). To characterize all 14 retained pairs, we measure changes from the last pre-administration observation, standardize them by the channel's median absolute change between consecutive training measurements, and subtract reference drift matched by channel, elapsed time, and starting-value decile. We report the largest adjusted deviation in the label-supported direction as $|\mathrm{ES}|$, with a dash indicating no peak in that direction. These summaries define the descriptive groups in Table 5, but neither affect probe inclusion nor establish an untreated counterfactual. For the five DAT probes, we also derive query windows from these curves by selecting the peak bin in the supported direction and extending to adjacent bins with the same sign and at least half the peak magnitude.

We query at the geometric midpoint $\omega_{\mathrm{mid}} = \sqrt{\omega_{\mathrm{lo}}\omega_{\mathrm{hi}}}$, a representative time on a multiplicative scale. All times are measured from administration onset, and windows specify evaluation horizons rather than the full duration of drug action. For each instance, we use the recorded ground-truth history ending immediately before the target drug's first administration, condition on the current intervention step, and construct an inquiry for the target channel at ω_{mid} hours after administration. No post-administration measurements or subsequent interventions are included. Removal deletes only the target drug's order lines at the current step, keeping other interventions and all earlier history fixed. All arms share the target channel and query time. Since evaluation compares predictions across arms, no recorded measurement is required at the exact query time. Negative controls use channel-specific windows, so comparisons with positive probes reflect their respective query horizons.

Each instance is evaluated at the midpoint under every available edit condition, so these counts refer to instances rather than model queries. Except for USR, which uses aggregate positive and control sensitivities, metrics are computed per probe and then macro-averaged. Thus, *Score* is averaged after its per-probe calculation, and *Rev* includes only probes with an opposing arm.

C More Experimental Details

C.1 More details of MedRoAM.

Illustration of the differentiable numerical surrogate. In §3.3, we construct a differentiable numerical surrogate by weighting each numeric token's value according to its decimal position in the observed response. For brevity, let $\mu_k^b = \sum_{u \in \mathcal{T}_{\text{num}}} n(u) P_k^b(u)$, so that $v_{\text{sur}}^b = \eta \sum_{k=1}^{K} \pi_k \mu_k^b$. The following examples illustrate this construction under different tokenizations.

Case 1. Single-digit tokens. Suppose the observed response is 1250, tokenized as `[1, 2, 5, 0]`. The corresponding place values are $(1000, 100, 10, 1)$, and the numerical surrogate is defined as:

$$v_{\text{sur}}^b = 1000\mu_1^b + 100\mu_2^b + 10\mu_3^b + \mu_4^b.$$

For a tokenizer whose numeric tokens are exactly the ten digits, $\mu_k^b = \sum_{d=0}^{9} d P_k^b(d)$, recovering digit-wise weighting. When each distribution assigns probability one to the observed token, the surrogate recovers $1000 \times 1 + 100 \times 2 + 10 \times 5 + 0 = 1250$.

Case 2. Multi-digit tokens. Suppose the same response is tokenized as `[125, 0]`. The token `125` has numerical value 125 and place value 10, since its rightmost digit occupies the tens position. Thus, the surrogate can be computed as:

$$v_{\text{sur}}^b = 10\mu_1^b + \mu_2^b.$$

Concentrating the probabilities on the observed tokens again recovers $10 \times 125 + 0 = 1250$. To illustrate a soft prediction, suppose $P_1^b(\texttt{125}) = 0.8$, $P_1^b(\texttt{138}) = 0.2$, and $P_2^b(\texttt{0}) = 1$. In this case, $v_{\text{sur}}^b = 10(0.8 \times 125 + 0.2 \times 138) = 1276$. Thus, the surrogate varies continuously with token probabilities without selecting a discrete output.

Case 3. Decimal values. Suppose the observed response is 0.85, tokenized as `[0, ., 85]`. The decimal point determines the place values but contributes no numeric-token term. The two numeric tokens have place values 1 and 10^{-2}, respectively, yielding:

$$v_{\text{sur}}^b = \mu_1^b + 10^{-2}\mu_2^b.$$

Concentrating the probabilities on `0` and `85` recovers $0 + 10^{-2} \times 85 = 0.85$. For a negative observed response, the fixed sign $\eta = -1$ negates the weighted sum.

In all cases, the observed response determines the token positions, place values, and sign, which are shared across the three intervention arms. Only the model probabilities vary with the intervention. Notably, these probabilities come from the full-vocabulary softmax without renormalization over numeric tokens, so nonnumeric tokens contribute zero. Since the probabilities are differentiable functions of the model logits, the surrogate supports gradient-based alignment training. However, it represents a numerical value under a fixed observed-response format rather than the expected value of unrestricted generation. Alternative tokens may have different digit lengths, while the assigned place values remain fixed.

Implementation details. PSD uses token-position correspondences between original and self-generated versions of the same serialized trajectory. When a generated value changes length, the positions of later value boundaries are adjusted by the resulting token offset. Thus, each pair refers to the same inquiry occurrence, including repeated measurements of a channel. Only the first-token distributions at these matched value boundaries are distilled. For IRA, we select the first observed measurement of the target channel within the intervention's response window and use the same observed value string for all three arms. This training response need not occur at the window midpoint. Evaluation instead constructs a query at the geometric midpoint without requiring an observed measurement there. IRA uses a two-level reduction. Eligible instance losses are first averaged within each example, and the resulting losses are then averaged over examples assigned recorded histories in the microbatch. A microbatch without eligible IRA instances contributes zero

IRA loss. This can occur even though every selected trajectory contains a probe, because the history mixture may assign every example a rollout history. The channel scale Δ_q has a lower bound of 1e-6 solely to prevent division by zero. This safeguard is inactive for the channel scales in our data.

C.2 Implementation Details of Baselines

C.2.1 LLMs and medical LWMs

Cutting-edge LLMs. This baseline tests whether pretrained LLMs can perform clinical world modeling without task-specific training. We keep model weights unchanged and evaluate all LLM baselines using the same trajectory-continuation protocol as the LWMs. In practice, each model receives the serialized trajectory prefix ending immediately before the target inquiry value and continues it to generate a numerical prediction. Numerical values can be parsed directly from nearly all outputs. For the few remaining cases, we use regular expressions to extract the numerical value from the continuation. Under teacher-forced rollout, predictions are conditioned on the ground-truth history. Under autoregressive rollout, predicted values replace the corresponding observations in the trajectory text, and status tags are updated accordingly before subsequent predictions. Both settings retain the recorded intervention sequence and measurement schedule, differing in whether historical inquiry values are observed or predicted.

EHRWorld. We adopt EHRWorld (Mu et al., 2026) as the base medical LWM underlying our framework. To ensure consistent training data and model backbones, we reproduce EHRWorld on the ICU corpus described in §2.2 using Qwen3-4B/8B/14B. We follow the original training recipe and hyperparameter settings, including the serialization of trajectories into interleaved inquiry and intervention events. Intervention events provide conditioning information and are represented with empty outcomes. The state-prediction objective in Eq. 2 supervises only inquiry value tokens, excluding names, timestamps, units, and other non-value fields from the loss. The resulting checkpoints provide a common initialization for both EHRWorld (align) and MedRoAM, allowing us to evaluate further improvements upon EHRWorld through standard continued training and our proposed MedRoAM.

EHRWorld (align). To account for the benefits of additional training on the selected data, we continue training the initial EHRWorld model on the same training subset used for MedRoAM, while retaining the original EHRWorld training objective. This baseline provides a direct comparison between standard continued training and refinement with our PSD and IRA objectives. Comparing MedRoAM with both EHRWorld and EHRWorld (align) helps distinguish its overall improvements over the initial model from the additional gains beyond continued training on the selected trajectories.

C.2.2 Latent World Models

Shared structured representation. The two latent WM baselines, CWM (Wu et al., 2026) and ODE-RNN (Rubanova et al., 2019), operate on a structured dataset recovered from the same ICU trajectories used in our LWMs. Each stay is represented as a variable-length sequence $\{(\mathbf{s}_t, \mathbf{a}_t, \Delta t_t)\}_{t=1}^{T}$, where the state $\mathbf{s}_t \in \mathbb{R}^{25}$ covers vital signs, urine output, and laboratory measurements, aligned one-to-one with the inquiry channels of our models. The action $\mathbf{a}_t$ encodes the full drug vocabulary of the training set (430 drugs) as a multi-hot administration indicator, together with the accumulated log-dosage within the step. The term Δt_t denotes the elapsed time in hours between the t-th and $(t{+}1)$-th observations, which we keep at its true value rather than discretizing it into fixed bins, so that the irregular sampling of ICU records is preserved. At each observation step, missing channels are forward-filled from their most recent recorded values, with training-set medians used when no prior observation is available. A binary mask identifies genuinely observed channels for loss computation. States are z-score normalized per channel using training-set statistics, which are reused at test time along with the drug vocabulary. The data splits follow those of the main experiments.

CWM. Clinical World Model (CWM) (Wu et al., 2026) represents a patient's latent physiological state with a recurrent hidden state and predicts action-conditioned transitions. At each step, the current state, previous action, and elapsed time are projected and fed into a two-layer GRU. A transition head then predicts the *residual* change from the hidden state and current action, which is added to the current state to obtain the next-state prediction. Thus, the model predicts changes relative to the current state rather than absolute next-state values. Training minimizes a masked mean squared error over observed transitions, with only measured channels contributing to the loss. The original design

also includes a mortality prediction head, which we disable in the reported checkpoints to compare world models under the same transition objective. The model contains approximately 1M parameters.

ODE-RNN. To test whether a continuous-time inductive bias helps on irregularly sampled ICU trajectories, we implement the ODE-RNN of Rubanova et al. (2019). The hidden state evolves autonomously between observations according to a neural ordinary differential equation, and is updated discretely by a GRU cell whenever an observation arrives, which lets the model condition on the true elapsed time rather than on step indices alone. For batched training, we approximate the integral with a fixed number of Euler steps over the largest interval in the batch, and obtain each sample's hidden state by linearly interpolating according to its own Δt, trading exactness for parallelism. The transition head and residual parameterization follow the CWM, and the objective is the same masked transition MSE. The model contains roughly 0.7M parameters.

Inference protocol. Both latent WM baselines are evaluated under teacher-forced and autoregressive rollout settings. Under teacher-forced rollout, each prediction is conditioned on the recorded state $\mathbf{s}_t$, with unmeasured channels filled using their most recent recorded values when available, and their training-set medians otherwise. Under autoregressive rollout, the model is initialized with $\mathbf{s}_0$, the state at the trajectory's first observation step. Channels measured at this step use their recorded values, while unmeasured channels are filled with their training-set medians. The model then recursively feeds back its predictions, using the recorded actions and elapsed times between successive observation steps. Interventions preceding the first observation step are not included in this action sequence. Prediction and scoring begin at the second observation step, excluding $\mathbf{s}_0$ from evaluation. All metrics are computed after de-normalization in each channel's original units.

C.3 Training and Evaluation Details

To obtain the initial LWMs, we follow EHRWorld's training hyperparameters and perform full-parameter fine-tuning for three epochs with a peak learning rate of 1e-6, a cosine learning-rate schedule, a warm-up ratio of 0.05, and an effective batch size of 64. The maximum sequence length is 16,384 tokens, and the training uses bf16 precision and DeepSpeed ZeRO-3. Our implementation builds on `LLaMA-Factory`[3], extended with a custom data processor and trainer.

For MedRoam refinement training, we initialize the student $\mathcal{M}_\theta$ from the corresponding initial LWM checkpoint and update all parameters on the selected training subset for one epoch with a learning rate of 1e-7. EHRWorld (align) starts from the same checkpoint and undergoes full-parameter fine-tuning on the same subset using the standard LWM objective. Both methods use the same learning rate, effective batch size, number of epochs, and optimization steps, ensuring a controlled comparison under matched optimization steps. We select the default hyperparameters through evaluation on 200 development examples. Specifically, we set $\lambda = 0.1$ and $\beta = 0.3$. PSD uses a recorded-history sampling probability of $\rho = 0.5$ and reverse KL divergence. For IRA, we set the margin to $m = 2.0$ and the substitution tolerance to $\gamma = 0.5$, and cap the directional loss $\mathcal{L}_{\text{dir}}$ at $2m = 4.0$ to stabilize training. For the "-w/o PSD" ablation in §4.4, we replace $\mathcal{L}_{\text{PSD}}$ with $\mathcal{L}_{\text{LWM}}$ and use recorded histories throughout. We independently apply IRA to each example with probability 0.5, matching the expected IRA supervision frequency of the full model at $\rho = 0.5$. All experiments are conducted on eight NVIDIA H100 (80GB) GPUs.

At inference, we use greedy decoding for reproducibility, with `vLLM`[4] for offline batched generation. Value generation is capped at 40 tokens and stops at the line break, since each prediction is a single number. A value that cannot be parsed is counted as an error rather than discarded. For evaluation, following EHRWorld, we measure numerical accuracy using *Numer. S@25* and clinical status agreement using *Stat F1*. Let v_j and $\hat{v}_j$ denote the ground-truth and predicted values of the j-th evaluated measurement, and n_{eval} denote the total number of measurements. *Numer. S@25* measures the proportion of predictions whose absolute relative error does not exceed 25%:

$$\text{S@25} = \frac{1}{n_{\text{eval}}} \sum_{j=1}^{n_{\text{eval}}} \mathbf{1}\left\{ \frac{|\hat{v}_j - v_j|}{|v_j|} \leq 0.25 \right\}, \tag{6}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. All evaluated ground-truth values are nonzero, avoiding zero denominators. For *Stat F1*, let $[r_j^{\text{lo}}, r_j^{\text{hi}}]$ denote the clinical reference interval for measurement j. We

[3] https://github.com/hiyouga/LLaMA-Factory
[4] https://github.com/vllm-project/vllm

define its status mapping as $\mathrm{abn}_j(v) = \mathbf{1}\{v < r_j^{\mathrm{lo}} \vee v > r_j^{\mathrm{hi}}\}$, where 1 denotes abnormal and 0 denotes normal. Applying this mapping to v_j and $\hat{v}_j$ yields the ground-truth and predicted status labels. We compute *Stat F1* as the macro-average of the normal and abnormal F1 scores:

$$\text{Stat F1} = \frac{1}{2} \sum_{c \in \{0,1\}} \frac{2\,\mathrm{TP}_c}{2\,\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{7}$$

where TP_c, FP_c, and FN_c are computed by treating class c as positive and the other class as negative. Both classes receive equal weight in the final score. Note that both metrics are reported as percentages under teacher-forced ("Next") and autoregressive ("Full") rollouts. For each metric, we additionally report the retention rate as $\mathrm{Ret.}(\%) = 100 \times \mathrm{Full}/\mathrm{Next}$, which quantifies the proportion of teacher-forced performance retained during autoregressive rollout.

D More Experiments

Beyond the main experiments, we analyze the key components of MedRoam, PSD and IRA, and evaluate the method under additional settings. Notably, due to computational constraints, we focus on the 4B scale, comparing MedRoam-4B with EHRWorld-4B and EHRWorld-4B (align). Smaller-scale experiments with the 8B and 14B models show consistent trends, suggesting that the 4B results are representative. Moreover, all hyperparameters are selected using 200 development examples with the 4B model before evaluation on the held-out test set. The test-set hyperparameter results serve only as post-selection sensitivity analyses. We apply the same settings across models without scale-specific tuning to ensure that performance gains do not stem from additional tuning.

D.1 Analysis of PSD

Effect of different components. There are two key components in our PSD. Specifically, the first is "hybrid-input", which mixes ground-truth and cached rollout histories as student inputs. The second is "KL", which supervises the student using the teacher's soft token distributions. To examine their contributions, we evaluate two variants: (1) "-w/o hybrid-input," which uses only cached rollout histories while retaining KL supervision; and (2) "-w/o KL," which retains history mixing but replaces KL supervision with maximum-likelihood training on observed value tokens. To ensure fairness, for both types of history input, we compute the loss only at the first token of each corresponding value. Notably, to isolate these effects from IRA, we train PSD and its variants without using the IRA approach.

Table 9: **Ablation of PSD components.** We evaluate teacher-forced (Next) and autoregressive (Full) performance on the ID test set using the 4B model.

Method	Numer. S@25		Stat F1	
	Next	Full	Next	Full
Scheduled Sampling	85.40	72.98	82.75	73.48
PSD	**87.09**	**73.51**	**84.15**	74.42
-w/o hybrid-input	79.39	72.05	80.34	**74.91**
-w/o KL	84.75	72.94	82.94	73.84

We additionally compare PSD with Scheduled Sampling (Bengio et al., 2015), a standard approach to mitigating the mismatch between training and inference. Our implementation mixes ground-truth histories with model-generated rollout histories and gradually increases the proportion of rollout histories during training. The key distinction is the supervision, as Scheduled Sampling trains against ground-truth hard labels, whereas PSD uses the teacher's soft token distributions.

We evaluate teacher-forced and autoregressive performance on the ID test set using *Numer. S@25* and *Stat F1*. As shown in Table 9, removing either component degrades performance. In particular, "-w/o hybrid-input" substantially reduces all metrics except *Stat F1* (Full), suggesting that relying exclusively on cached rollout histories may exacerbate forgetting. Removing KL supervision also reduces performance, supporting the benefit of the teacher's soft token distributions over one-hot targets when learning from imperfect histories. Scheduled Sampling exhibits a related trade-off, as it improves autoregressive performance but compromises teacher-forced performance. In contrast, full PSD improves autoregressive performance while largely preserving teacher-forced accuracy. These results support the complementary roles of history mixing and KL supervision in PSD.

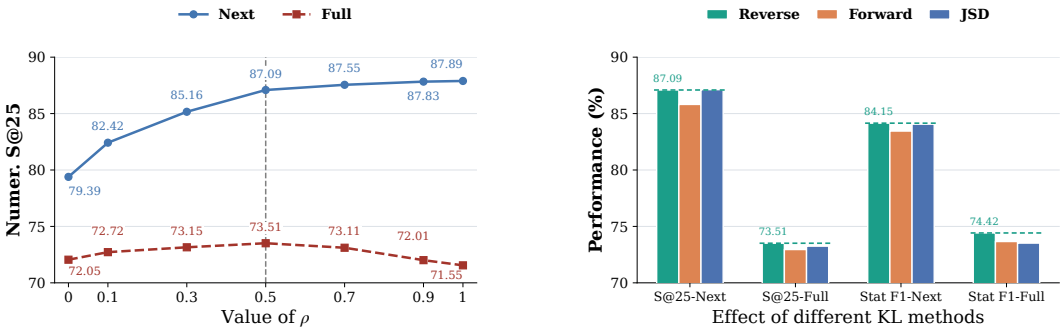


Figure 5: **Analysis of PSD. Left**: Teacher-forced (Next) and autoregressive (Full) performance under different mixing ratios ρ. **Right**: Performance comparison between reverse KL, forward KL, and JSD objectives. In both experiments, we evaluate our MEDROAM-4B (PSD-only) on the ID test set.

Effect of history mixing ratio ρ. To mitigate forgetting during PSD training, we mix ground-truth and cached rollout histories using a history mixing ratio ρ, where larger ρ assigns more ground-truth histories to the student. While exposure to cached rollouts may improve autoregressive robustness, ground-truth histories help preserve teacher-forced accuracy. To examine this trade-off, we vary ρ over $\{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$ and train with PSD alone to isolate its effects from IRA. Figure 5 (**Left**) reports the ID teacher-forced and autoregressive performance of MEDROAM-4B, measured by *Numer. S@25*. As ρ increases, greater exposure to ground-truth histories better preserves teacher-forced performance. However, a large ρ (*e.g.*, 0.9) limits exposure to cached rollouts, yielding smaller gains in autoregressive accuracy. Reducing ρ initially improves autoregressive performance, but excessively small values (*e.g.*, 0.1) degrade both teacher-forced and autoregressive accuracy. We hypothesize that self-distillation alone may provide insufficient guidance when student histories are dominated by rollout errors, while reduced exposure to ground-truth histories may exacerbate forgetting. Among the tested settings, $\rho = 0.5$ achieves the best autoregressive performance.

Effect of different divergence objectives. Model distillation can use different divergence objectives (Wu et al., 2025), including forward KL, reverse KL, and Jensen–Shannon divergence (JSD). To examine their effects, we train MEDROAM-4B with PSD alone and evaluate teacher-forced and autoregressive performance on the ID test set using *Numer. S@25* and *Stat F1*. As shown in Figure 5 (**Right**), reverse KL achieves competitive teacher-forced performance and the best autoregressive results on both metrics, while forward KL generally performs worse. We hypothesize that the mode-covering tendency of forward KL encourages the student to assign probability mass across the teacher's output distribution, which may be difficult when conditioning on noisy rollout histories. In contrast, the mode-seeking tendency of reverse KL may help the student concentrate on high-probability teacher outputs, making it better suited to precise numerical prediction (Gu et al., 2024). These results motivate our use of reverse KL for self-distillation in our PSD training.

D.2 ANALYSIS OF IRA

Effect of individual loss components. IRA constructs three intervention arms: the original intervention, its removal, and substitution with a similarly acting intervention. Two loss terms, $\mathcal{L}_{\text{dir}}$ and $\mathcal{L}_{\text{equiv}}$, encourage pharmacologically consistent response differences under removal and limit discrepancies under substitution, respectively. To examine their contributions, we remove each term from the full MEDROAM objective and train the corresponding 4B variants. We also include "-w/o both", which uses PSD alone. The models are evaluated on ID autoregressive rollout accuracy and held-in intervention sensitivity using *Numer. S@25* and *Sens×Mono*, respectively. Table 10 presents the comparative results, and we reveal two findings. First, removing $\mathcal{L}_{\text{equiv}}$ increases intervention sensitivity but reduces rollout accuracy. Without the equivalence constraint, the model may learn larger response differences under intervention removal, while becoming less constrained in its predictions, thus damaging the rollout robustness. Second, removing $\mathcal{L}_{\text{dir}}$ largely preserves rollout accuracy but substantially weakens intervention sensitivity, suggesting that explicit directional supervision is important for learning pharmacologically consistent responses. In general, these results highlight the complementary roles of the two loss terms, as their combination provides a better balance between rollout accuracy and intervention sensitivity.

Table 10: **Ablation of IRA loss components.** We remove the equivalence loss $\mathcal{L}_{\text{equiv}}$, directional loss $\mathcal{L}_{\text{dir}}$, or both (*i.e.*, using PSD alone). Subscripts indicate absolute changes relative to the full MEDROAM-4B model.

Method	Numer. S@25	Sens × Mono
MEDROAM-4B (full)	73.36	48.11
-w/o $\mathcal{L}_{\text{equiv}}$	$72.82_{\downarrow 0.54}$	$50.28_{\uparrow 2.17}$
-w/o $\mathcal{L}_{\text{dir}}$	$73.32_{\downarrow 0.04}$	$27.23_{\downarrow 20.88}$
-w/o both	$73.51_{\uparrow 0.15}$	$28.43_{\downarrow 19.68}$

Table 11: **Ablation of IRA stabilization strategies.** We remove ground-truth-only IRA supervision (clean-only), directional loss clipping (clip), or both. Subscripts indicate absolute changes relative to the full MEDROAM-4B model.

Method	Numer. S@25	Sens × Mono
MEDROAM-4B (full)	73.36	48.11
-w/o clean-only	$72.88_{\downarrow 0.48}$	$51.64_{\uparrow 3.53}$
-w/o clip	$72.76_{\downarrow 0.60}$	$41.01_{\downarrow 7.10}$
-w/o both	$71.87_{\downarrow 1.49}$	$41.30_{\downarrow 6.81}$

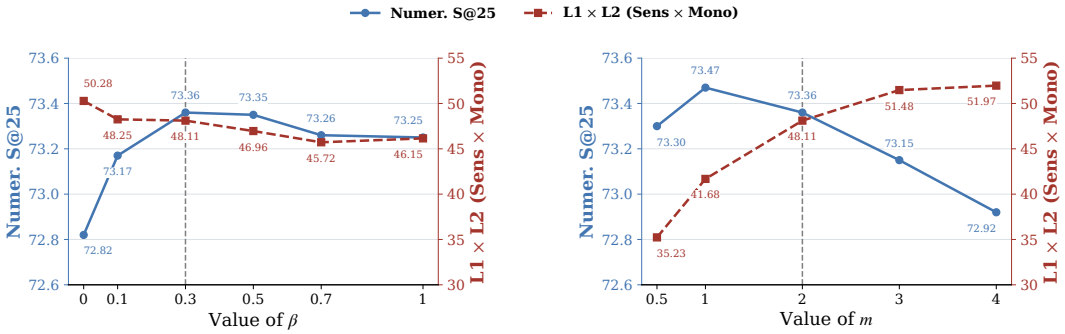


Figure 6: **Parameter analysis of IRA.** We vary the equivalence loss weight β with $m = 2$ (**Left**) and the directional margin m with $\beta = 0.3$ (**Right**). *Numer. S@25* measures autoregressive rollout accuracy on the ID test set, while *Sens* × *Mono* measures held-in intervention sensitivity.

Stabilizing IRA training. In preliminary experiments, we observed instability in vanilla IRA training, including loss spikes exceeding 70 and poor performance on both rollout accuracy and intervention sensitivity. Therefore, we introduce two stabilization strategies: (1) "clean-only", which applies IRA only when the student receives ground-truth histories; and (2) "clip", which caps $\mathcal{L}_{\text{dir}}$ at $2m$. To assess their contributions, we remove each strategy individually ("-w/o clean-only" and "-w/o clip") and both together ("-w/o both"), evaluating ID autoregressive rollout accuracy and L1–L2 held-in probing performance using *Numer. S@25* and *Sens*×*Mono*, respectively. As shown in Table 11, removing both strategies reduces the two metrics by 1.49 and 6.81 points, respectively. Either strategy improves rollout accuracy over this variant, while their combination provides the best balance between accuracy and intervention sensitivity. Notably, "-w/o clean-only" increases intervention sensitivity at the expense of rollout accuracy. This variant applies IRA to both recorded and rollout histories, increasing supervision exposure while also changing the history distribution. Thus, this comparison does not isolate the effect of history quality from that of supervision exposure. In contrast, "-w/o clip" degrades both metrics, with intervention sensitivity falling slightly below that of "-w/o both". We hypothesize that uncapped loss spikes can produce excessively large updates that disrupt learned predictions, highlighting the importance of clipping for stable IRA optimization.

Parameter analysis. There are two key hyperparameters in IRA: the factor β, which controls the weight of $\mathcal{L}_{\text{equiv}}$, and the margin m, which specifies the desired response separation between the recorded and removal arms in $\mathcal{L}_{\text{dir}}$. We examine their effects by varying one parameter while fixing the other at its default value. All experiments use MEDROAM-4B and evaluate ID autoregressive rollout accuracy and held-in L1–L2 probing tasks using *Numer. S@25* and *Sens*×*Mono*, respectively. First, with $m = 2$, we vary β over $\{0, 0.1, 0.3, 0.5, 0.7, 1\}$. Figure 6 (**Left**) shows that intervention sensitivity generally decreases as β increases, consistent with the loss-component ablation. A stronger equivalence constraint may limit the model's response variation, reducing sensitivity while helping preserve rollout accuracy. Rollout accuracy peaks at $\beta = 0.3$ and remains relatively stable thereafter, decreasing by only 0.11 points at $\beta = 1$. We hypothesize that an excessively large β increases the relative contribution of IRA to the joint objective, potentially interfering with PSD optimization. Second, with $\beta = 0.3$, we vary m over $\{0.5, 1, 2, 3, 4\}$. As shown in Figure 6 (**Right**), intervention sensitivity increases with m, whereas rollout accuracy peaks at $m = 1$ and then declines. A larger margin encourages greater response differences between intervention arms, which may strengthen

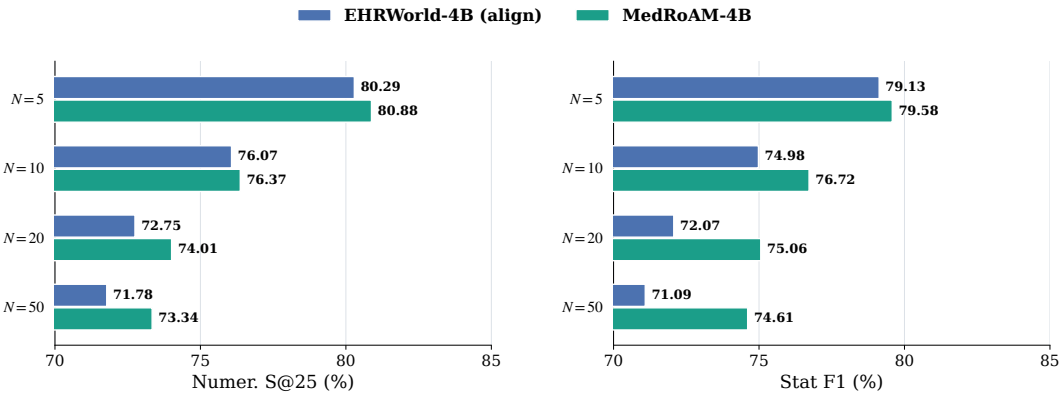


Figure 7: **Performance comparison across autoregressive rollout horizons.** We compare EHRWorld-4B (align) and MEDROAM-4B when predicting the next N patient states from ground-truth histories, measured by *Numer. S@25* (left) and *Stat F1* (right) on the ID test data.

Table 12: **Held-in intervention probing under teacher-forced and autoregressive histories.** We evaluate the 4B models using ground-truth histories and their own autoregressive histories with $N = 5$. Additionally, we evaluate EHRWorld-4B (align) on histories generated by MEDROAM-4B.

Method	History	*Intervention Removal* (L1–L2)			*Dose* (L3)	*Intervention Specificity* (L4)		
		Sens (↑)	Mono (↑)	Score (↑)	Dose-ρ (↑)	USR (↓)	SER (↑)	Rev (↑)
EHRWorld-4B (align)	Ground-truth	48.29	67.85	32.09	0.096	53.31	18.88	27.22
EHRWorld-4B (align)	Own rollout ($N = 5$)	45.12	64.85	28.88	0.076	53.88	17.32	23.29
EHRWorld-4B (align)	MEDROAM rollout ($N = 5$)	44.74	68.53	29.63	0.071	51.16	17.66	26.49
MEDROAM-4B	Ground-truth	**60.85**	78.80	**48.11**	0.165	**44.91**	**31.46**	**52.99**
MEDROAM-4B	Own rollout ($N = 5$)	58.86	**80.79**	47.92	**0.186**	45.14	28.47	52.79

sensitivity at the expense of preserving accurate trajectory predictions. More specifically, in the case of $m = 2$, MEDROAM offers a favorable trade-off, as it can better balance improved intervention sensitivity with a modest reduction in rollout accuracy.

D.3 MORE IN-DEPTH ANALYSES

How does MEDROAM perform across different rollout horizons? In our main experiments, we follow EHRWorld and evaluate autoregressive rollouts over complete trajectories. Here, we further examine performance at different rollout horizons. Specifically, we vary N over $\{5, 10, 20, 50\}$ and autoregressively predict the next N patient states conditioned on observed ground-truth histories. We compare EHRWorld-4B (align) and MEDROAM-4B using *Numer. S@25* and *Stat F1* on the ID test data. As shown in Figure 7, both models exhibit declining performance as N increases, consistent with error accumulation over longer rollouts. Nevertheless, MEDROAM-4B consistently outperforms EHRWorld-4B (align) across all horizons, with larger gains at longer horizons, particularly in *Stat F1*. These results support the role of PSD in mitigating the training–inference mismatch and improving robustness to accumulated rollout errors.

Does MEDROAM retain intervention sensitivity under autoregressive histories? Our main probing experiments use teacher-forced ground-truth histories. However, in practice, models must respond appropriately to interventions when conditioning on their own generated histories. To examine whether MEDROAM retains this capability, we autoregressively generate the five history steps immediately preceding each probing step ($N = 5$) and repeat the evaluation using these histories. To assess the influence of history quality, we additionally evaluate EHRWorld-4B (align) on histories generated by MEDROAM-4B, denoted as "MEDROAM rollout".

Table 12 shows the results on the L1-L4 held-in probing tasks, from which we reveal two findings. First, autoregressive histories modestly reduce several probing scores relative to ground-truth histories. Nevertheless, MEDROAM retains a substantial advantage over EHRWorld-4B (align) across L1–L4. Second, conditioning EHRWorld-4B (align) on MEDROAM rollouts improves most probing metrics

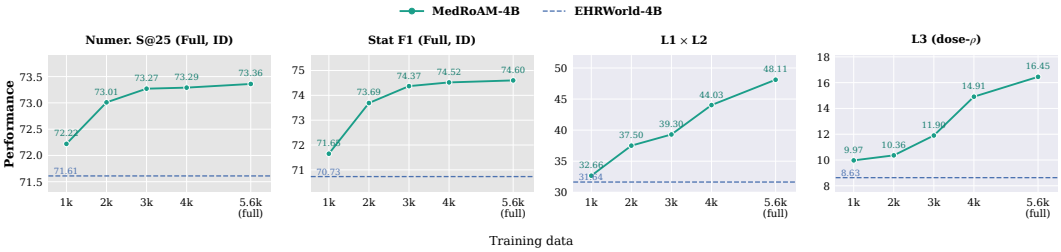


Figure 8: **Effect of training data size.** We train MEDROAM-4B on subsets of the full 5.6k training set and evaluate autoregressive rollout accuracy on the ID test set (first two panels) and held-in intervention sensitivity (last two panels). Dose-ρ is scaled by 100 for display.

Table 13: **State prediction performance (%) across additional LLM backbones.** We evaluate teacher-forced ("Next") and autoregressive ("Full") rollouts on ID and OOD test sets.

Method	Numer. S@25 (ID)			Stat F1 (ID)			Numer. S@25 (OOD)			Stat F1 (OOD)		
	Next	Full	Ret.	Next	Full	Ret.	Next	Full	Ret.	Next	Full	Ret.
(a) MedGemma1.5-4B												
EHRWorld	87.56	73.17	83.57	83.74	72.17	86.18	80.46	68.55	85.20	80.17	72.00	89.81
EHRWorld (align)	**87.57**	73.41	83.83	**83.77**	72.38	86.40	**80.60**	68.51	85.00	**80.22**	72.21	90.01
MEDROAM (Ours)	**87.57**	**73.79**	**84.26**	83.74	**73.79**	**88.12**	79.78	**68.56**	**85.94**	80.02	**73.27**	**91.56**
(b) Ministral3-8B												
EHRWorld	88.18	73.98	83.90	**84.27**	72.81	86.40	**81.18**	68.12	83.91	**80.60**	69.07	85.69
EHRWorld (align)	**88.20**	74.04	83.95	84.25	72.89	86.52	81.13	67.90	83.69	80.59	68.69	85.23
MEDROAM (Ours)	88.08	**74.74**	**84.85**	84.23	**74.32**	**88.23**	80.04	**68.74**	**85.88**	80.34	**71.14**	**88.55**
(c) MiniCPM5-2B												
EHRWorld	**87.49**	71.82	82.09	**83.96**	70.27	83.69	**80.15**	66.01	82.36	**80.58**	69.97	86.83
EHRWorld (align)	87.44	71.69	81.99	83.88	69.82	83.24	**80.15**	66.02	82.37	80.51	69.58	86.42
MEDROAM (Ours)	87.39	**72.93**	**83.45**	83.88	**72.85**	**86.85**	80.07	**66.46**	**83.00**	80.33	**72.11**	**89.77**

compared with its own rollouts, although *Sens* and *Dose-ρ* decline slightly. We hypothesize that more accurate generated patient states contribute to these improvements. However, even with the same MEDROAM-generated histories, EHRWorld-4B (align) remains substantially less sensitive to interventions than MEDROAM. Together, these results suggest that the intervention sensitivity learned by MEDROAM extends to autoregressive histories and support the complementary benefits of rollout robustness and intervention alignment.

How does training data size affect the performance of MEDROAM? To examine the effect of training data size, we randomly sample subsets of 1k, 2k, 3k, and 4k trajectories from the full 5.6k training set to train MEDROAM-4B. We evaluate autoregressive rollout accuracy on the ID test set using *Numer. S@25* and *Stat F1*, and intervention sensitivity on held-in L1–L3 probing tasks using *Sens*$\times$*Mono* and *Dose-ρ*. For comparison, we include the initially trained EHRWorld-4B as a baseline. As shown in Figure 8, performance improves consistently across all four metrics as training data size increases. Notably, even with only 1k training trajectories, MEDROAM-4B outperforms the baseline on all evaluated metrics. These results support the data efficiency of our method and show that additional training data benefits both rollout accuracy and intervention sensitivity.

Does MEDROAM generalize to other LLM backbones? Our main experiments demonstrate that MEDROAM improves rollout robustness and intervention sensitivity with Qwen3 backbones. To examine its applicability to other LLMs, we evaluate three additional backbones, MedGemma1.5-4B[5], Ministral3-8B[6], and MiniCPM5-2B[7], representing medical, general-purpose, and edge-oriented LLMs, respectively. First, we train each backbone on the full training corpus using the same hyperparameters in Appendix C.3, obtaining initial medical LWMs denoted as EHRWorld. We then refine these models with MEDROAM using the default hyperparameters and cached rollouts generated by each corresponding backbone. For comparison, we continue standard LWM training on the curated

[5] https://huggingface.co/google/medgemma-1.5-4b-it
[6] https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512
[7] https://huggingface.co/openbmb/MiniCPM5-2B

Table 14: **Intervention probing across additional LLM backbones.** The tasks assess removal responsiveness (L1), directional consistency (L2), dose sensitivity (L3), and intervention specificity (L4). "H-In" and "H-Out" denote held-in and held-out intervention-response pairs, respectively.

Method	*Intervention Removal* (L1–L2)						*Dose* (L3)		*Intervention Specificity* (L4)					
	Sens (%, ↑)		Mono (%, ↑)		Score (%, ↑)		Dose-ρ (↑)		USR (%, ↓)		SER (%, ↑)		Rev (%, ↑)	
	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out
(a) MedGemma1.5-4B														
EHRWorld	46.22	55.51	64.16	75.80	29.36	44.67	0.476	0.092	57.12	**38.43**	19.92	**17.44**	27.90	**47.50**
EHRWorld (align)	47.12	56.45	66.98	75.48	30.86	45.61	0.492	0.082	51.02	39.90	21.32	16.76	29.21	46.37
MEDROAM (Ours)	**56.07**	**58.10**	**71.36**	**78.26**	**40.36**	**48.57**	**0.527**	**0.174**	**45.36**	43.30	**24.52**	15.84	**41.99**	46.63
(b) Ministral3-8B														
EHRWorld	43.32	54.92	60.78	**71.55**	25.57	41.68	0.497	0.332	68.11	**43.89**	18.33	20.60	24.85	41.66
EHRWorld (align)	45.28	54.36	64.23	69.66	28.16	40.86	0.517	0.356	60.38	47.43	18.18	22.63	27.96	41.62
MEDROAM (Ours)	**56.33**	**58.93**	**70.03**	69.73	**40.21**	**44.37**	**0.556**	**0.373**	**51.67**	48.64	**25.96**	**22.98**	**41.52**	**44.41**
(c) MiniCPM5-2B														
EHRWorld	36.33	39.58	62.31	71.68	21.65	30.56	0.382	0.146	62.65	51.58	12.43	11.39	27.12	32.39
EHRWorld (align)	36.53	38.20	63.52	73.31	22.07	29.99	0.383	0.127	64.02	52.19	12.84	13.50	26.12	31.61
MEDROAM (Ours)	**52.25**	**52.53**	**79.59**	**75.88**	**41.71**	**41.99**	**0.479**	**0.174**	**51.18**	**51.02**	**21.50**	**16.96**	**59.51**	**49.52**

Table 15: **Paired bootstrap comparison of MedRoAM and EHRWorld (align) across backbone models.** Entries report the difference (MedRoAM - EHRWorld (align)) with 95% confidence intervals from patient-clustered paired bootstrap resampling ($B = 2000$). *Numer. S@25* and *Stat F1* are evaluated on ID Full and reported in percentage points, while *Sens* (L1), *Mono* (L2), and *Sens* × *Mono* are evaluated on the 14 held-in probes. * indicates that the confidence interval excludes zero.

Backbone	*Numer. S@25*	*Stat F1*	*Sens* (L1)	*Mono* (L2)	*Sens* × *Mono*
Qwen3-4B	$+1.61$ $[+1.36, +1.88]^*$	$+3.53$ $[+3.22, +3.84]^*$	$+12.56$ $[+10.14, +15.05]^*$	$+10.95$ $[+8.12, +13.56]^*$	$+16.02$ $[+13.69, +18.21]^*$
Qwen3-8B	$+1.83$ $[+1.51, +2.05]^*$	$+3.55$ $[+3.25, +3.85]^*$	$+15.33$ $[+13.08, +17.55]^*$	$+10.10$ $[+7.24, +12.92]^*$	$+19.14$ $[+16.98, +21.17]^*$
Qwen3-14B	$+1.60$ $[+1.34, +1.88]^*$	$+3.53$ $[+3.25, +3.82]^*$	$+12.41$ $[+10.06, +14.71]^*$	$+12.35$ $[+9.57, +15.03]^*$	$+18.46$ $[+16.27, +20.61]^*$
MedGemma1.5-4B	$+0.38$ $[+0.17, +0.56]^*$	$+1.41$ $[+1.22, +1.63]^*$	$+8.95$ $[+6.66, +11.15]^*$	$+4.38$ $[+1.64, +7.03]^*$	$+9.50$ $[+7.49, +11.48]^*$
Ministral3-8B	$+0.70$ $[+0.51, +0.89]^*$	$+1.43$ $[+1.20, +1.67]^*$	$+11.05$ $[+8.78, +13.23]^*$	$+5.80$ $[+2.83, +8.93]^*$	$+12.05$ $[+9.98, +14.22]^*$
MiniCPM5-2B	$+1.24$ $[+1.02, +1.45]^*$	$+3.03$ $[+2.76, +3.29]^*$	$+15.72$ $[+13.33, +18.08]^*$	$+16.07$ $[+13.15, +18.90]^*$	$+19.64$ $[+17.37, +21.76]^*$

subset to obtain EHRWorld (align). Following the main experiments, we evaluate teacher-forced and autoregressive performance on the ID and OOD test sets, together with intervention sensitivity on the 14 held-in and 7 held-out pairs through L1–L4 probing. As shown in Table 13, MEDROAM improves autoregressive rollout performance over both EHRWorld and EHRWorld (align), narrowing the gap between teacher-forced and autoregressive prediction. Furthermore, Table 14 shows improvements in intervention sensitivity, with MEDROAM achieving the best results on most metrics across the evaluated backbones. These gains extend from the 14 pairs used for IRA supervision to the seven held-out intervention-response pairs, although some held-out metrics show mixed results. Overall, these findings suggest that the benefits of MEDROAM extend beyond Qwen3 to MedGemma, Ministral, and MiniCPM, supporting its applicability across different LLM backbones.

Bootstrap Confidence Intervals. To assess the robustness of MEDROAM's gains over EHRWorld (align), we conduct patient-level paired bootstrap resampling (Efron & Tibshirani, 1994) with 2,000 replicates. Specifically, we evaluate *Numer. S@25* and *Stat F1* under autoregressive rollout on the ID test set, together with *Sens* (L1), *Mono* (L2), and *Sens* × *Mono* on the 14 held-in probes. For each backbone, we sample patients with replacement while retaining their associated observations. Both methods are evaluated on the same resampled patients, and we compute the performance difference as MEDROAM minus EHRWorld (align). Table 15 reports the estimated differences and their 95% confidence intervals across six backbones. All intervals lie entirely above zero, supporting positive improvements in every comparison. The gains range from 0.38 to 1.83 points for *Numer. S@25* and from 1.41 to 3.55 points for *Stat F1*. On the probing set, gains range from 8.95 to 15.72 points for *Sens*, from 4.38 to 16.07 points for *Mono*, and from 9.50 to 19.64 points for *Sens* × *Mono*. For example, on Qwen3-4B, the *Numer. S@25* gain is 1.61 points with a 95% confidence interval of [1.36, 1.88], while the *Sens* × *Mono* gain is 16.02 points with an interval of [13.69, 18.21]. These results support the robustness of the observed improvements to patient-sampling variability.

Consistency across random seeds. We assess training-seed variability using the 4B model and three matched refinement seeds for EHRWorld (align) and MEDROAM, while keeping the initial LWM checkpoint fixed. Figure 9 presents the individual results, means, and standard deviations across the three seeds. We evaluate *Numer. S@25* and *Stat F1* under ID Full autoregressive rollout, together with *Sens*×*Mono* on held-in and held-out probes. Across these runs, MEDROAM shows

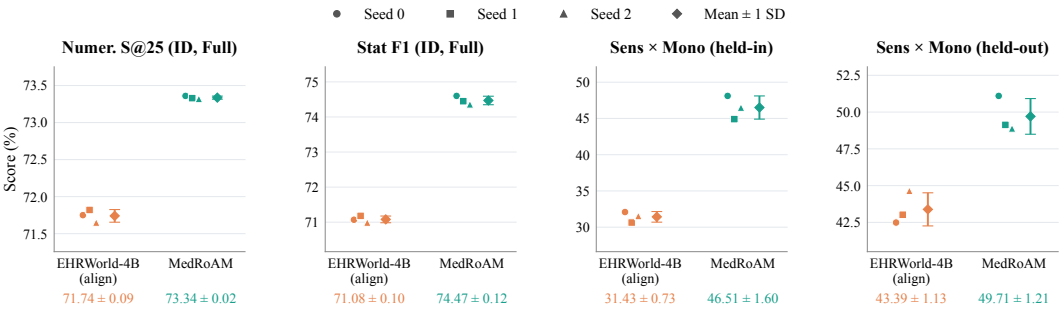


Figure 9: **Consistency across random seeds.** We compare EHRWorld-4B (align) and MEDROAM using three matched training random seeds. Error bars indicate ± 1 standard deviation.

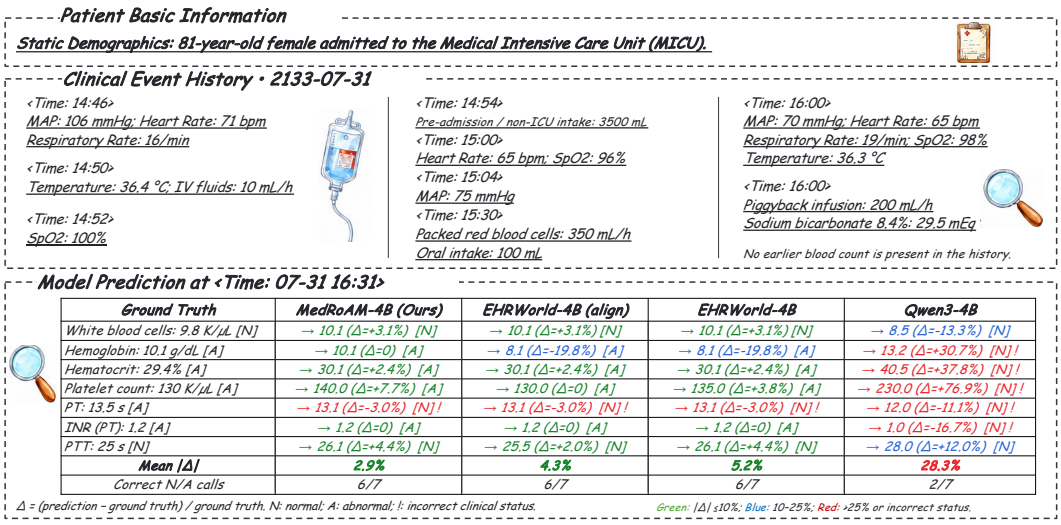


Figure 10: **Qualitative comparison under teacher-forced rollout.** Four models predict the same next observation for an 81-year-old female MICU patient. Notably, colors indicate absolute relative error: green ($|\Delta| \leq 10\%$), blue ($10\% < |\Delta| \leq 25\%$), and red ($|\Delta| > 25\%$ or an incorrect normal/abnormal classification), where $\Delta = (\hat{y} - y)/y$. "N" and "A" denote normal and abnormal status, respectively, and "!" marks a misclassification.

slightly higher variability in *Stat F1* and both probing scores, but lower variability in *Numer. S@25*. We conjecture that random sampling between ground-truth and cached rollout histories in PSD may contribute to this variability. Nevertheless, MEDROAM outperforms EHRWorld (align) on all four metrics in every matched-seed comparison. For instance, across the three seeds, the averaged gains are 1.60 points for *Numer. S@25*, 3.39 for *Stat F1*, 15.08 for the held-in probing score, and 6.32 for the held-out probing score. Notably, these averages differ from the single-run gains reported in the main tables, and the main tables report results from refinement seed 0. Overall, these results support the consistency of the observed gains across the evaluated refinement seeds.

Case Study. To have a close look, we compare MEDROAM-4B with EHRWorld-4B (align), EHRWorld-4B, and Qwen3-4B on two patient cases under teacher-forced and autoregressive rollout. Figure 10 presents the teacher-forced case for an 81-year-old female MICU patient. The queried step contains the first blood count of the ICU stay, with no earlier hemoglobin, hematocrit, or platelet measurements in the history. The models must draw on the available clinical context, including packed red blood cells administered at 350 mL/h approximately one hour earlier, sodium bicarbonate, and intravenous fluids. Five of the seven target measurements are abnormal. Qwen3-4B misses all five abnormalities, predicting normal-range hemoglobin, hematocrit, platelets, PT, and INR, and correctly classifies only 2/7 measurements. In contrast, MEDROAM-4B and both EHRWorld variants identify four of the five abnormalities, achieving 6/7 correct classifications and missing only the borderline PT value. Despite their identical classification accuracy, their numerical predictions differ. MEDROAM-4B predicts hemoglobin exactly at 10.1 g/dL, whereas both EHRWorld variants predict

Patient Basic Information

Static Demographics: 57-year-old male admitted to the Coronary Care Unit (CCU).

Previous 5 Steps · 2180-10-10

‹Time: 07:00› MAP: 80 mmHg; Heart Rate: 98 bpm; Respiratory Rate: 16/min; SpO_2: 94%

Time / Measurement	Ground Truth	MedRoAM-4B (Ours)	EHRWorld-4B (align)	EHRWorld-4B	Qwen3-4B
08:45 / Glucose	109 [A]	→ 120 [A]	→ 103 [N] !	→ 100 [N] !	→ 140 [A]
08:45 / Lactate	2 [N]	→ 1.1 [N]	→ 1.1 [N]	→ 1.1 [N]	→ 1.8 [N]
09:25 / Glucose	165 [A]	→ 142 [A]	→ 103 [N] !	→ 100 [N] !	→ 140 [A]
10:14 / Glucose	178 [A]	→ 160 [A]	→ 104 [N] !	→ 102 [N] !	→ 140 [A]
10:32 / Glucose	181 [A]	→ 162 [A]	→ 105 [N] !	→ 105 [N] !	→ 140 [A]

Each prediction is fed back into that model's subsequent context. N: normal; A: abnormal; !: incorrect clinical status.

Autoregressive Model Prediction at ‹Time: 10-10 11:20›

Ground Truth	MedRoAM-4B (Ours)	EHRWorld-4B (align)	EHRWorld-4B	Qwen3-4B
Glucose: 181 [A]	→ 160 (Δ=-11.6%) [A]	→ 105 (Δ=-42.0%) [N] !	→ 105 (Δ=-42.0%) [N] !	→ 140 (Δ=-22.7%) [A]
White blood cells: 12.8 [A]	→ 12.8 (Δ=0) [A]	→ 10.1 (Δ=-21.1%) [N] !	→ 10.1 (Δ=-21.1%) [N] !	→ 10.2 (Δ=-20.3%) [N] !
Hemoglobin: 9.9 [A]	→ 8.8 (Δ=-11.1%) [A]	→ 11.1 (Δ=+12.1%) [A]	→ 11.1 (Δ=+12.1%) [A]	→ 13.8 (Δ=+39.4%) [A]
Hematocrit: 28.6 [A]	→ 25.4 (Δ=-11.2%) [A]	→ 32.3 (Δ=+12.9%) [A]	→ 32.3 (Δ=+12.9%) [A]	→ 41 (Δ=+43.4%) [N] !
Platelet count: 95 [A]	→ 140 (Δ=+47.4%) [A]	→ 180 (Δ=+89.5%) [N] !	→ 171 (Δ=+80.0%) [N] !	→ 230 (Δ=+142.1%) [N] !
PT: 16.8 [A]	→ 15.7 (Δ=-6.5%) [A]	→ 12.4 (Δ=-26.2%) [N] !	→ 12.3 (Δ=-26.8%) [N] !	→ 12 (Δ=-28.6%) [N] !
INR (PT): 1.5 [A]	→ 1.4 (Δ=-6.7%) [A]	→ 1.16 (Δ=-22.7%) [A]	→ 1.15 (Δ=-23.3%) [A]	→ 1 (Δ=-33.3%) [N] !
PTT: 32.3 [N]	→ 31.1 (Δ=-3.7%) [N]	→ 27.6 (Δ=-14.6%) [N]	→ 28.7 (Δ=-11.1%) [N]	→ 30 (Δ=-7.1%) [N]
Lactate: 3.2 [A]	→ 2.7 (Δ=-15.6%) [A]	→ 1.1 (Δ=-65.6%) [N] !	→ 1.1 (Δ=-65.6%) [N] !	→ 1.8 (Δ=-43.8%) [N] !
Mean |Δ|	12.6%	34.1%	32.8%	42.3%
Correct N/A calls	9/9	4/9	4/9	3/9

Δ = (prediction − ground truth) / ground truth. Clinical status follows the reference ranges supplied with the data.

Green: \|Δ\| ≤10%; Blue: 10–25%; Red: >25% or incorrect status.

Figure 11: **Qualitative comparison under autoregressive rollout.** We compare predictions for a 57-year-old male CCU patient. Each model feeds its own predicted values back into the subsequent context from admission. The upper section shows the five preceding steps, and the lower section compares predictions for the queried step. Colors and clinical-status markers follow Figure 10.

8.1 g/dL, an underestimation of 19.8%. EHRWorld-4B (align) predicts the platelet count exactly at 130 K/μL, while MEDROAM-4B and EHRWorld-4B predict 140 and 135 K/μL, respectively. Overall, MEDROAM-4B achieves the lowest mean absolute relative error of 2.9%, compared with 4.3% for EHRWorld-4B (align), 5.2% for EHRWorld-4B, and 28.3% for Qwen3-4B.

Moreover, Figure 11 examines a separate 57-year-old male CCU patient under autoregressive rollout, where each model conditions on its own predictions from admission, with 65 values preceding the queried step. The preceding observations reveal a persistent difference in the predicted glucose trajectory. While the observed glucose rises from 109 to 165, 178, and 181, EHRWorld-4B (align) predicts 103, 103, 104, and 105, and EHRWorld-4B predicts 100, 100, 102, and 105, consistently misclassifying these measurements as normal. Qwen3-4B repeatedly predicts 140. In contrast, MEDROAM-4B predicts 120, 142, 160, and 162, following the upward trend while retaining the abnormal classification. At the queried step, MEDROAM-4B correctly classifies all nine measurements, including eight abnormalities, compared with 4/9 for both EHRWorld variants and 3/9 for Qwen3-4B. Its mean absolute relative error is 12.6%, versus 34.1% for EHRWorld-4B (align), 32.8% for EHRWorld-4B, and 42.3% for Qwen3-4B. In particular, MEDROAM-4B predicts the white blood cell count exactly at 12.8 and identifies elevated lactate with a prediction of 2.7 versus the observed 3.2. All three baselines misclassify both measurements as normal. Nevertheless, its platelet prediction of 140 remains 47.4% above the observed value of 95, despite correctly identifying thrombocytopenia. Overall, these cases illustrate the potential of MEDROAM to track evolving measurements and preserve abnormal clinical states under autoregressive feedback, while highlighting that correct status classification does not necessarily imply precise numerical prediction.

D.4 DISCUSSION

Does directly adding pharmacological knowledge to the prompt improve intervention sensitivity? In MEDROAM, IRA explicitly trains the model to align its responses with pharmacological knowledge. To examine whether the same knowledge can improve predictions through prompting alone, we construct a baseline that includes the 14 curated intervention–response pairs in the model's context. Specifically, we describe each pair's expected direction, similar and opposing interventions, and response window in text. We apply this knowledge augmentation to EHRWorld and the "-w/o IRA" variant of MEDROAM, then evaluate both on held-in and held-out probing tasks.

Table 16 presents the results. For EHRWorld, knowledge prompting improves the held-in composite score at all three scales, although the gains are small at 4B and 14B. However, held-out results are mixed, with the composite score decreasing at 4B and 14B and improving only slightly at 8B.

Table 16: **Effect of pharmacological knowledge prompting on intervention probing.** "+ Knowledge" adds knowledge about the 14 held-in intervention-response pairs to the preceding model's prompt. "H-In" and "H-Out" denote held-in and held-out evaluation, respectively.

Method	*Intervention Removal* (L1–L2)						*Dose* (L3)		*Intervention Specificity* (L4)					
	Sens (%, ↑)		Mono (%, ↑)		Score (%, ↑)		Dose-ρ (↑)		USR (%, ↓)		SER (%, ↑)		Rev (%, ↑)	
	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out	H-In	H-Out
EHRWorld-4B	48.37	53.91	66.44	75.58	31.64	42.73	0.086	0.269	51.56	**42.14**	19.11	14.33	25.87	45.95
+ Knowledge	47.76	51.33	67.16	72.05	31.65	39.70	0.154	0.222	51.80	45.00	19.22	14.17	28.08	40.75
MEDROAM-4B -w/o IRA	52.86	56.53	58.05	72.05	28.43	45.12	0.035	0.238	46.73	51.30	22.13	17.81	21.24	32.21
+ Knowledge	52.81	53.68	58.48	74.54	29.60	42.20	0.105	0.192	55.60	54.60	23.07	18.35	23.62	29.32
MEDROAM-4B (Ours)	**60.85**	**61.02**	**78.80**	**79.89**	**48.11**	**51.10**	**0.165**	**0.281**	**44.91**	49.97	**31.46**	**20.59**	**52.99**	**56.85**
EHRWorld-8B	46.07	53.15	65.93	74.20	29.71	41.54	0.056	0.274	53.18	**40.02**	16.18	17.50	29.77	47.02
+ Knowledge	52.46	55.53	64.83	73.35	34.08	42.03	0.052	0.264	51.40	41.80	21.93	18.88	35.69	47.90
MEDROAM-8B -w/o IRA	51.96	61.78	58.01	73.00	29.16	47.57	−0.002	0.301	53.10	51.80	22.73	18.36	25.64	40.05
+ Knowledge	56.85	57.61	56.39	72.78	32.08	44.38	−0.019	0.313	51.00	56.80	23.92	19.14	30.15	36.31
MEDROAM-8B (Ours)	**62.47**	**65.08**	**76.21**	**76.47**	**49.68**	**52.22**	**0.185**	**0.361**	**42.51**	50.24	**31.80**	**20.70**	**57.39**	**62.66**
EHRWorld-14B	45.50	56.20	65.69	76.87	28.69	45.89	0.273	0.525	54.01	39.13	22.67	18.68	35.40	53.89
+ Knowledge	45.34	54.48	67.32	79.18	29.44	44.82	0.280	0.477	55.00	**38.20**	21.55	19.27	34.08	34.94
MEDROAM-14B -w/o IRA	48.32	57.85	54.47	79.19	24.71	47.49	0.052	0.501	53.80	48.60	21.58	**19.79**	23.44	34.94
+ Knowledge	47.18	55.69	55.47	78.15	24.40	45.29	0.132	0.503	54.50	47.20	23.67	17.80	23.09	36.81
MEDROAM-14B (Ours)	**59.02**	**58.26**	**78.04**	**82.14**	**48.01**	**50.74**	**0.370**	**0.597**	**41.33**	39.97	**30.51**	19.62	**58.75**	**65.44**

Similarly, the "-w/o IRA" variant shows inconsistent held-in gains, while its held-out composite score decreases at every scale. These results suggest that providing pharmacological knowledge in context does not ensure effective application or transfer to unseen relations. The decline on held-out probes may reflect distraction from knowledge that is less relevant to those probes. In contrast, the full MEDROAM achieves the highest composite scores on both held-in and held-out probes across all model scales. In general, these comparisons support the value of learning pharmacological response constraints through IRA rather than relying on prompting alone.

Agreement between the IRA surrogate and greedy-decoded numerical predictions. As described in §3.3, IRA uses a differentiable numerical surrogate to constrain intervention responses. However, changes in the surrogate do not guarantee changes in the generated value, since shifting probability mass among non-argmax numerical tokens can alter the surrogate without changing the selected token. Moreover, the surrogate uses ground-truth numerical prefixes and place weights determined by the observed value, whereas greedy decoding conditions on the model's preceding outputs and may produce a different number of digits or decimal places. To examine this discrepancy, we evaluate all probes on the training set using full-vocabulary greedy decoding, starting immediately before the first value token in each arm and parsing the generated output into a numerical value.

For each probe, we use the surrogate effect $d_{\text{or}} = (v^{\text{o}}_{\text{sur}} - v^{\text{r}}_{\text{sur}})/\Delta_q$ defined in §3.3 and compute the decoded effect as $d^{\text{dec}}_{\text{or}} = (v^{\text{o}}_{\text{dec}} - v^{\text{r}}_{\text{dec}})/\Delta_q$. Here, $v^{\text{o}}_{\text{dec}}$ and $v^{\text{r}}_{\text{dec}}$ are the values generated under the original and removal arms, respectively, and both effects use the same channel-specific scale Δ_q. Given the supported direction $\sigma_p \in \{-1, +1\}$, we define surrogate margin satisfaction as $H = \{\sigma_p d_{\text{or}} \geq m\}$ with $m = 2$. We define six metrics to quantify the agreement. *Satisfied* is $P(H)$, the proportion of instances satisfying the surrogate margin. *Unchanged*|*Satisfied* is $P(d^{\text{dec}}_{\text{or}} = 0 \mid H)$, the proportion of satisfied instances with identical decoded values. *Direction*|*Changed* is $P(\sigma_p d^{\text{dec}}_{\text{or}} > 0 \mid d^{\text{dec}}_{\text{or}} \neq 0)$, measuring directional consistency among nonzero decoded effects. *Spear-*ρ measures the Spearman rank correlation between d_{or} and $d^{\text{dec}}_{\text{or}}$ across all instances. *Changed* is $P(d^{\text{dec}}_{\text{or}} \neq 0)$, the proportion of instances with different decoded values. Finally, *Direction* is $P(\sigma_p d^{\text{dec}}_{\text{or}} > 0)$, measuring directional consistency across all instances, with unchanged values counted as failures.

Table 17 compares EHRWorld (align) with MEDROAM under different λ. We report results averaged over all instances rather than macro-averaged across probes. The $\lambda = 0$ setting provides a control without IRA training. Its *Changed* rate reaches 55.08%, compared with 36.39% for EHRWorld (align). However, only 30.06% of these changes follow the supported direction, and Spearman ρ is slightly negative at -0.068. Thus, more frequent numerical changes alone do not establish the desired directional response. At the default $\lambda = 0.1$, *Direction*|*Changed* reaches 84.19%, while *Direction* increases from 16.56% at $\lambda = 0$ to 64.72%. Moreover, Spearman ρ rises from -0.068 to 0.426, and *Unchanged*|*Satisfied* decreases from 37.62% to 8.25%, compared with 35.84% for EHRWorld (align). These results indicate stronger agreement between surrogate effects and decoded responses, although surrogate margin satisfaction does not guarantee a change in the generated value. Across λ, the metrics *Direction*, *Direction*|*Changed*, and Spearman ρ increase monotonically with

Table 17: **IRA surrogate–decoding agreement.** We evaluate all intervention-response probes on the training set using full-vocabulary greedy autoregressive decoding. "Satisfied" denotes surrogate margin satisfaction ($m = 2$), "Changed" denotes unequal decoded values across arms, and "Direction" denotes agreement with the expected pharmacological direction. "Spear-ρ" is the Spearman correlation between signed surrogate and decoded effects. All entries except Spear-ρ are percentages.

Setting	Satisfied	Unchanged|Satisfied ↓	Direction|Changed ↑	Spear-ρ ↑	Changed	Direction ↑
EHRWorld-4B (align)	3.35	35.84	41.93	−0.062	36.39	15.26
MEDROAM-4B						
$\lambda = 0$	3.92	37.62	30.06	−0.068	55.08	16.56
$\lambda = 0.05$	8.78	23.40	76.12	0.229	69.43	52.85
$\lambda = 0.1$	24.66	8.25	84.19	0.426	76.87	64.72
$\lambda = 0.2$	36.51	5.84	87.56	0.544	79.93	69.99
$\lambda = 0.5$	43.41	4.91	89.16	0.622	81.19	72.39
$\lambda = 1.0$	51.05	4.06	91.22	0.710	83.73	76.39

Table 18: **Absolute and relative sensitivity to negative controls.** Specifically, $\overline{\text{Sens}}_{\text{pos}}$ and $\overline{\text{Sens}}_{\text{ctrl}}$ denote independently macro-averaged sensitivities to positive and negative control probes, respectively. *USR* denotes their ratio, *i.e.*, $\overline{\text{Sens}}_{\text{ctrl}}/\overline{\text{Sens}}_{\text{pos}}$. All values are reported as percentages.

Method	Held-in Probing			Held-out Probing		
	$\overline{\text{Sens}}_{\text{pos}}$	$\overline{\text{Sens}}_{\text{ctrl}}$	USR	$\overline{\text{Sens}}_{\text{pos}}$	$\overline{\text{Sens}}_{\text{ctrl}}$	USR
EHRWorld-4B	48.37	24.94	51.56	53.91	22.72	42.14
EHRWorld-4B (align)	48.29	25.74	53.31	54.14	23.58	43.56
MEDROAM-4B	60.85	27.33	44.91	61.02	30.49	49.97
EHRWorld-8B	46.07	24.50	53.18	53.15	21.27	40.02
EHRWorld-8B (align)	47.14	24.41	51.79	53.33	20.60	38.62
MEDROAM-8B	62.47	26.56	42.51	65.08	32.70	50.24
EHRWorld-14B	45.50	24.57	54.01	56.20	21.99	39.13
EHRWorld-14B (align)	46.61	23.27	49.92	56.14	23.10	41.15
MEDROAM-14B	59.02	24.39	41.33	58.26	23.29	39.97

λ. Meanwhile, *Unchanged|Satisfied* decreases monotonically, reaching 4.06% at $\lambda = 1.0$, where *Direction* reaches 76.39% and Spearman ρ reaches 0.710. Notably, the conditional unchanged rate is computed only over probes satisfying the surrogate margin, whose proportion also increases with λ, rather than over the full evaluation set. However, larger IRA weights reduce rollout accuracy (Figure 4), indicating a trade-off between stronger intervention alignment and state prediction.

Overall, IRA uses probability-weighted numerical aggregation to make response constraints differentiable. This surrogate operates under a fixed observed-response format and does not represent the expected value of unrestricted numerical generation. Consequently, satisfying its margin does not always produce a corresponding change in the decoded value. The improvements in decoded directional consistency and effect correlation support the practical value of this approximation. Nevertheless, the remaining discrepancy is a limitation of IRA, and future work will explore closer agreement between surrogate and decoded numerical responses.

Absolute sensitivity to negative controls. In L4 probing, we use $\text{USR} = \overline{\text{Sens}}_{\text{ctrl}}/\overline{\text{Sens}}_{\text{pos}}$ to quantify responsiveness to unrelated-intervention removal relative to target-intervention removal, independently macro-averaging control and positive-probe sensitivities. A lower USR indicates lower relative responsiveness to controls, but does not necessarily imply lower absolute control sensitivity. To examine this distinction, Table 18 compares both sensitivities and USR of different LWM training methods across Qwen3 model scales on held-in and held-out probes. For held-in probes, MEDROAM reduces USR relative to both baselines, while absolute control sensitivity increases at every scale compared with EHRWorld (align). Thus, the lower ratio reflects a larger proportional increase in positive-probe sensitivity rather than reduced responsiveness to unrelated interventions. On held-out probes, control sensitivity also increases relative to both baselines, with particularly large increases at 4B and 8B. We hypothesize that this partly reflects the absence of explicit constraints on unrelated interventions during IRA training, which may make the model overly responsive to intervention changes. These findings highlight a limitation of our method, and future work will explore incorporating such constraints to improve intervention specificity.

Direct comparison of model responses to intervention removal and substitution. In the L4 probing task, we use SER to evaluate model responses after replacing a probing intervention with a pharmacologically similar intervention. We define the normalized decoded removal and substitution effects as $d_{\text{or}}^{\text{dec}} = (v_{\text{dec}}^{\text{o}} - v_{\text{dec}}^{\text{r}})/\Delta_q$ and $d_{\text{so}}^{\text{dec}} = (v_{\text{dec}}^{\text{s}} - v_{\text{dec}}^{\text{o}})/\Delta_q$. SER measures, within each probe, the percentage of instances satisfying $0 < |d_{\text{so}}^{\text{dec}}| < |d_{\text{or}}^{\text{dec}}|$ among those with valid predictions in all three arms and $|d_{\text{or}}^{\text{dec}}| > 0$, followed by macro-averaging across probes. This criterion requires a nonzero substitution effect smaller than the removal effect. However, a large removal effect allows a substantial substitution effect to satisfy the criterion. SER alone does not quantify the absolute magnitude of substitution effects. To examine these effects directly, we report the macro averages of $|d_{\text{so}}^{\text{dec}}|$ and $|d_{\text{or}}^{\text{dec}}|$ on the same channel-normalized scale. Unlike the sample-level averages in Figure 1 (Right), these macro averages give equal weight to each intervention-response pair. Hence, this difference in aggregation leads to different reported values.

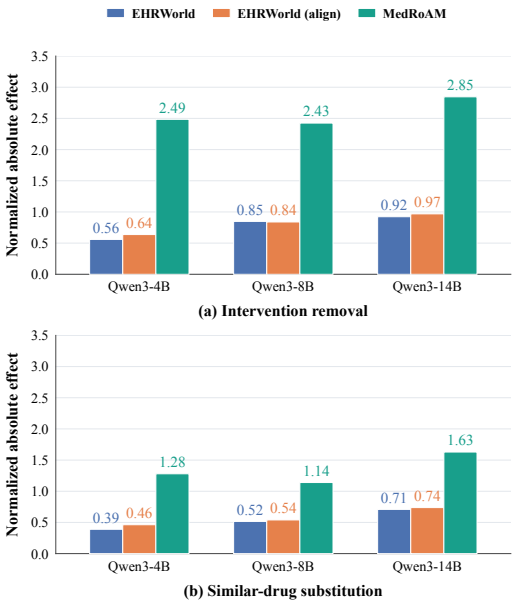


Figure 12: **Numerical responses to intervention changes.** (**a**) and (**b**) show the macro-averaged normalized absolute effects of intervention removal and similar-drug substitution.

Figure 12 presents the results on the held-in pairs across Qwen3 model sizes. Compared with EHRWorld and EHRWorld (align), MEDROAM produces larger removal effects at every scale. Its mean removal effects are 2.49, 2.43, and 2.85 for 4B, 8B, and 14B, respectively. The corresponding values for EHRWorld (align) are 0.64, 0.84, and 0.97. These results indicate greater sensitivity to intervention removal. The mean substitution effects of MEDROAM increase to 1.28, 1.14, and 1.63, compared with 0.46, 0.54, and 0.74 for EHRWorld (align). However, the substitution effects remain smaller than the removal effects on average. The ratios of mean substitution to mean removal effects are 0.52, 0.47, and 0.57 for MEDROAM, below the corresponding ratios of 0.73, 0.64, and 0.76 for EHRWorld (align). These aggregate ratios describe relative effect magnitudes and are distinct from SER. The nonzero substitution effects are consistent with the design of $\mathcal{L}_{\text{equiv}}$. The tolerance parameter $0 < \gamma < 1$ permits substitution effects within a bound determined by the removal effect and the margin floor. The objective encourages similar responses rather than identical predictions. This flexibility accommodates possible differences in patient responses to pharmacologically similar drugs and different doses. Conversely, EHRWorld and EHRWorld (align) produce smaller effects under both substitution and removal. Their smaller substitution effects may partly reflect limited sensitivity to intervention changes rather than a better understanding of pharmacological equivalence.

E SCOPE AND LIMITATIONS

Scope. Our main intervention evaluation holds out patient episodes while reusing the 14 supervised relations. Moreover, we evaluate seven held-out intervention-response pairs excluded from IRA supervision, providing evidence of generalization beyond the supervised pairs. However, these evaluations do not establish generalization to entirely unseen drugs or pharmacological mechanisms. Together with dose variation and opposing substitutions, these tests characterize pharmacological consistency under specified inputs rather than patient-specific causal effects.

Limitations. Our work has three main limitations. First, IRA relies on an indirect numerical surrogate and provides incomplete supervision of intervention specificity. Specifically, its teacher-forced numerical surrogate uses the observed sign, token layout, place values, and response prefixes, so its influence on freely decoded values is indirect. That is, satisfying the surrogate constraints does not guarantee corresponding changes in decoded predictions. Moreover, IRA supervises intervention removal and similar substitutions without explicitly constraining responses to unrelated or opposing interventions. Although opposing-response scores improve empirically, absolute sensitivity to

unrelated interventions can also increase, indicating room for improvement in intervention specificity. Future work will explore closer surrogate-decoding agreement and broader intervention supervision. Second, although our refinement experiments cover Qwen3, MedGemma, Ministral, and MiniCPM backbones, they remain limited to models with up to 14B parameters. Evaluating larger models and a broader range of backbones would further clarify the method's scalability and general applicability. Finally, we evaluate the models' capabilities without directly assessing their utility for downstream clinical decision-making. Future work will examine whether the observed improvements translate into better performance on counterfactual reasoning tasks.