MEDAGENTWORLD: PRIVILEGED ON-POLICY SELF-DISTILLATION FOR WORLD-MODEL-AUGMENTED CLINICAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical large language models perform well on static case questions but are rarely evaluated on the sequential decisions of acute care. We present MedAgentWorld (MedWA), a world-model-augmented reinforcement learning framework for multi-turn clinical planning. We curate MedWA-118K and MedWA-Bench as time-ordered patient trajectories and build MedWA-Gym, where an agent acts on currently available information and receives replayed observations. Because historical records lack responses to valid requests without recorded responses, MedWA-WM predicts missing next observations. We adapt on-policy self-distillation into privileged OPSD: a fixed teacher receives the target observation, subsequent examination results, and final diagnosis, and guides the student through token-level Jensen–Shannon divergence on student-generated response prefixes. At inference, the student conditions only on the visible patient history and selected action. We train MedWA-Agent with supervised fine-tuning (SFT) and reinforcement learning (RL) on replayed and world-model-generated branches. On MedWA-Bench, the agent reaches 44.92% diagnosis accuracy, compared with 42.73% without the world model, and MedWA-WM improves held-out next-observation prediction over prompted language-model baselines. To probe generalization beyond MC-MED, we further evaluate MedWA-WM on an out-of-distribution MIMIC-IV-ED test set converted to the same interaction schema, where it continues to outperform prompted baselines across all reported metrics. These results support world-model-assisted learning for interactive clinical planning.

1 INTRODUCTION

Medical large language models (LLMs) have shown strong potential in healthcare applications. Recent systems demonstrate strong performance in medical question answering, case understanding, clinical knowledge assessment, and diagnosis prediction (Singhal et al., 2023; 2025; McDuff et al., 2025). However, most existing work still treats clinical intelligence as a static prediction problem: a model receives a completed case summary or fixed findings and directly outputs a diagnosis or recommendation. This setup is convenient for benchmarking, but it compresses real care into a one-shot input-output mapping. In practice, clinicians operate under incomplete information, repeatedly asking, testing, treating, waiting, and revising plans as new evidence arrives. Clinical intelligence needs to move beyond static prediction toward multi-turn clinical planning.

In clinical care, this gap is particularly salient: diagnostic workup and therapeutic management are tightly coupled, and clinical decisions entail not only selecting interventions but also deferring or withholding them. Yet most current benchmarks do not capture the temporal structure of care: what information is available at each step, which actions are taken, what responses follow, and which alternatives are postponed. Real clinical records naturally contain time-ordered tests, interventions, returned results, vital-sign changes, and clinical events (Kansal et al., 2025). Motivated by this difference, we construct **MedWA-Bench** and curate **MedWA-118K**, representing patient care as time-ordered patient trajectories and moving evaluation from static prediction to trajectory-level assessment.

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.

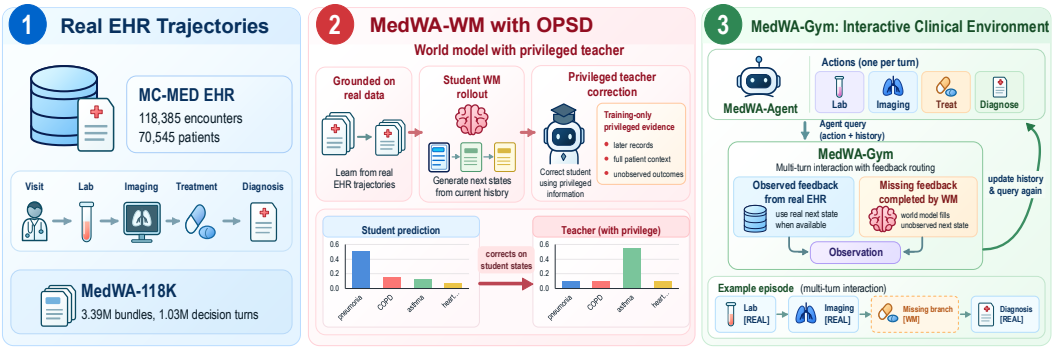

Figure 1: **MedAgentWorld overview. (1) Trajectories:** Electronic health record (EHR) encounters become time-ordered patient histories. **(2) Privileged on-policy self-distillation (OPSD):** A student world model generates observations; a fixed teacher uses the recorded response, later examination results, and final diagnosis to guide it through JSD on student-generated prefixes. Only the student runs at inference. **(3) Multi-turn MedWA-Gym:** The agent chooses an action, receives recorded feedback when available or a world-model response otherwise, updates its history, and acts again until diagnosis. The bottom strip illustrates this loop.

Trajectory-level data are not enough if the learning problem is still reduced to predicting the next recorded action. A clinical agent should instead act under partial information, choose the next diagnostic or therapeutic step, and receive the resulting patient feedback over multiple rounds. To support this setting, we build **MedWA-Gym**, a rollout environment for clinical interaction. In MedWA-Gym, the agent can observe only the patient information revealed so far and must choose the next action from the current history. The environment then returns the next patient feedback, allowing the agent to learn when to request evidence and when to commit to a diagnosis.

A key limitation remains: historical trajectories cover only the actions that were actually taken. If the agent selects a schema-valid test, imaging study, or intervention that does not appear in the observed trajectory, the environment cannot directly return the corresponding real response. The resulting requests fall outside the actions covered by replay, leaving the agent without feedback for exploring unobserved decision branches. We refer to these as counterfactual branches; their responses must be estimated from the available evidence.

To address this problem, we introduce **MedWA-WM**, a response world model that predicts the next observation for schema-valid requests without recorded feedback. A second challenge is exposure to the model's own predictions during generation. We adapt on-policy self-distillation (Zhao et al., 2026) into **privileged OPSD** for clinical feedback. A fixed teacher receives the student's question and generated prefix, together with training-only hints: the recorded target observation, subsequent examination results, and final diagnosis. Token-level Jensen–Shannon divergence aligns the student with this informed teacher at the same prefixes. At inference, the student receives only the visible history and request. We then train **MedWA-Agent** in MedWA-Gym with both replayed observations and world-model-generated branches, forming **MedAgentWorld (MedWA)** for multi-turn clinical planning.

Our contributions are fourfold. First, we construct MedWA-Bench and MedWA-118K for trajectory-level evaluation beyond static prediction. Second, we introduce MedWA-Gym, which organizes patient trajectories into an interactive rollout environment. Third, we propose MedWA-WM and OPSD to predict missing responses using privileged supervision on student-generated response prefixes. Fourth, we train MedWA-Agent with replayed and imagined branches for multi-turn clinical planning. On MedWA-Bench, MedWA-Agent improves diagnosis accuracy and action-bundle matching over the same agent trained without world-model feedback, and MedWA-WM further transfers to an out-of-distribution MIMIC-IV-ED evaluation converted to the same interaction schema, where it retains the lead over prompted language-model baselines.

2 Related Work

Medical LLMs. Medical LLMs mainly answer questions and predict diagnoses from completed cases (Singhal et al., 2023; 2025; McDuff et al., 2025; Tang et al., 2024). Interactive work adds history taking (Tu et al., 2025b), sequential test acquisition (Nori et al., 2025), and virtual EHR settings (Jiang et al., 2025; Schmidgall et al., 2024). MedWA evaluates the broader sequence of evidence acquisition, intervention, waiting, and final diagnosis.

Clinical RL. Retrospective clinical RL has learned sepsis treatment policies (Komorowski et al., 2018; Yu et al., 2021) and studied offline safety (Nambiar et al., 2023; Wang et al., 2024; Tu et al., 2025a), often with narrow action spaces. Historical records lack feedback for unchosen actions, a limitation shared by language agents for evidence acquisition (Bani-Harouni et al., 2025). MedWA combines replay with modeled short-horizon feedback for these branches.

World Models. Model-based RL uses predicted dynamics to support policy learning (Ha & Schmidhuber, 2018; Hafner et al., 2019; Schrittwieser et al., 2020), including in interactive environments (Agarwal et al., 2025; Bruce et al., 2024; Parker-Holder & Fruchter, 2025; Bar et al., 2025; Feng et al., 2025). Medical world models learn EHR dynamics (Xu et al., 2026; Mu et al., 2026), while transformer methods estimate clinical counterfactual outcomes (Melnychuk et al., 2022; Rahman et al., 2023). Emergency datasets such as MIMIC-IV-ED support trajectory-based evaluation (Johnson et al., 2023). MedWA uses a text LLM to generate structured feedback across heterogeneous, unobserved actions while grounding interaction in real-trajectory replay.

3 Data Curation

Source and trajectory construction. We build MedWA-118K and MedWA-Bench from MC-MED's 118,385 emergency encounters involving 70,545 patients (Kansal et al., 2025). Records from visits, orders, laboratories, imaging, waveforms, and medical and medication histories are linked by encounter and ordered by clinical time. Each trajectory starts with information available at presentation: demographics, chief complaint, triage findings, vital signs, and background history. Laboratory and imaging results are revealed only after their associated requests. A bundle groups related orders into one requestable unit, such as a laboratory panel. Laboratory and imaging bundles provide structured feedback. The final diagnosis is the terminal label.

Corpus and benchmark. MedWA-118K contains the full trajectory corpus, averaging 28.6 requestable bundles (8.79 laboratory, 4.13 imaging, and 15.69 treatment) and 8.73 decision turns per encounter, for approximately 3.39 million bundles and 1.03 million turns. A decision turn is one agent action, which may request multiple bundles. MedWA-Bench retains cases with valid diagnoses and disease categories, then stratifies by disease and an activity-based complexity score. Its 220 cases average 17.63 requestable bundles, 8.31 feedback bundles, and 8.73 decision turns.

4 Method

We build the proposed framework in three components: **MedWA-Gym**, a rollout environment for sequential clinical interaction; **MedWA-WM**, a response world model for valid requests without recorded feedback; and **MedWA-Agent**, which is trained in the resulting MedWA-Gym hybrid environment for multi-turn clinical planning.

4.1 MedWA-Gym: Rollout Environment for Sequential Clinical Interaction

Sequential diagnostic setting. A rollout is a sequence of agent actions and environment responses within one encounter. In **MedWA-Gym**, the agent chooses diagnostic or therapeutic actions from the currently available history and receives feedback over multiple turns. Formally, we model the interaction as a multi-turn, partially observable decision process rather than a one-shot prediction problem. Let the interactive environment be denoted by

$$\mathcal{E} = (\mathcal{S}, \mathcal{H}, \mathcal{A}, \mathcal{O}, P, \rho_0, T_{\max}),$$

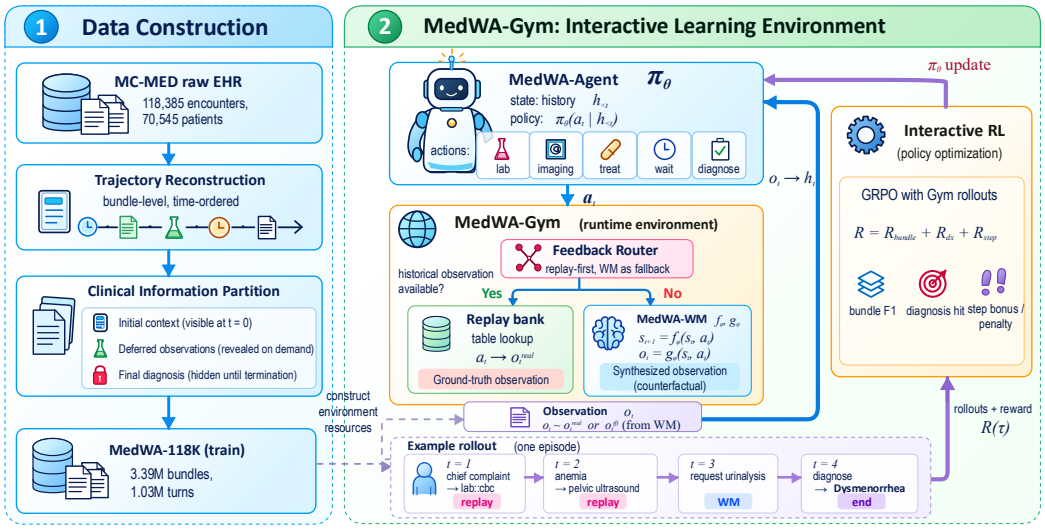

Figure 2: **MedAgentWorld overview. (1) Data:** MC-MED records become time-ordered trajectories for training and replay. **(2) MedWA-Gym:** observed action branches return real feedback; the OPSD-trained MedWA-WM fills missing observations. **(3) Agent training:** Supervised fine-tuning initializes the policy before reinforcement learning optimizes it on hybrid rollouts. The bottom strip shows a four-turn episode.

where $\mathcal{S}$ is the latent patient-state space, $\mathcal{H}$ is the space of observable histories, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, P is the environment transition rule, ρ_0 is the initial case distribution, and $T_{\max}$ is the turn limit. The latent state $s_t \in \mathcal{S}$ represents the patient's underlying physiology; the agent never observes s_t directly and acts only from the accumulated history $h_t \in \mathcal{H}$, the context available before action a_t.

Episodes and action encoding. Each encounter provides an initial context, requestable bundles, recorded observations, and a terminal diagnosis. Initial context contains presentation-time demographics, chief complaint, triage findings, vital signs, and background history; later results are revealed through interaction. The agent selects namespaced bundle identifiers, such as `lab::cbc_hematology`, from the case-specific requestable set. JSON calls encode one or more bundles, alongside control actions for waiting, diagnosis, and stopping. The 20 integer-coded action labels in the reference event-replay prototype are a separate, coarser interface (Appendix B).

Replay-grounded interaction. A request is valid when it belongs to the defined request schema and uses the required calling format. This check establishes interface validity, rather than patient-specific clinical appropriateness. For a valid feedback-bearing request a_t, observation routing is

$$o_t = \begin{cases} o_t^{\text{real}}, & \text{if } a_t \in \mathcal{D}_{\text{obs}}(h_t), \\ \hat{o}_t, & \text{if recorded feedback is unavailable,} \end{cases}$$

and the history is updated by appending the action and response: $h_{t+1} = (h_t, a_t, o_t)$. Here, $\mathcal{D}_{\text{obs}}(h_t)$ contains requests with recorded feedback available in the current episode. Multi-bundle actions are resolved at the bundle level; waiting and termination follow the environment rules. Recorded feedback takes priority. For valid requests without recorded responses, MedWA-WM supplies a prediction conditioned on the visible history. Requests outside the schema or with malformed calls fail validation.

4.2 MedWA-WM: Grounded World Modeling with Self-Distillation

Patient State and World Model. MedWA-WM uses Qwen3-4B (Yang et al., 2025) as its language-model backbone to predict structured feedback for requests without recorded responses. The latent state s_t denotes underlying patient physiology; the implemented model conditions on its observable

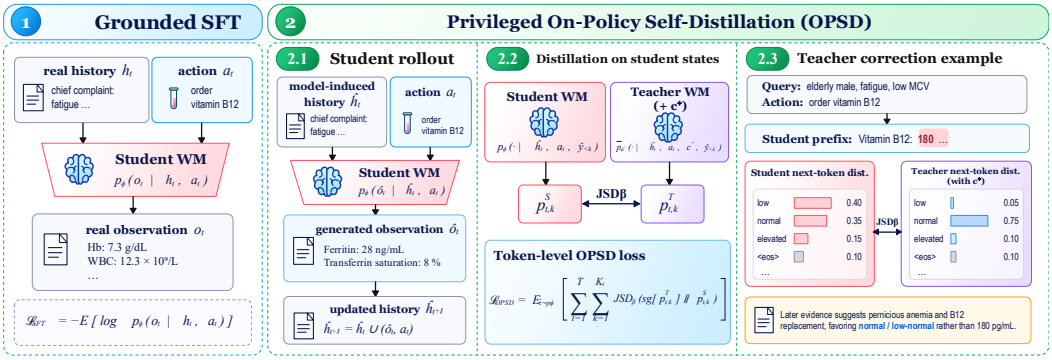

Figure 3: Illustration of **MedWA-WM** with privileged on-policy self-distillation (OPSD). After SFT, the student generates a response from the visible history and request. A fixed teacher also receives privileged evidence c_t^+ and predicts at the same student-generated prefixes. Token-level JSD updates only the student; hints are training-only.

history h_t, rather than accessing s_t directly. Diagnostic requests seek evidence about that state, while treatment history provides context for subsequent observations. Given action a_t, the response model is

$$p_\theta(o_t \mid h_t,\, a_t),$$

where θ denotes world-model parameters. Training examples follow the interaction interface and action format used by MedWA-Gym. During agent RL, MedWA-WM is served through vLLM to return structured feedback for the requested action.

Grounding on Observed EHR Transitions. We first ground the world model on observed EHR transitions (h_t, a_t, o_t) using supervised next-observation learning:

$$\mathcal{L}_{\mathrm{SFT}} = -\,\mathbb{E}[\log p_\theta(o_t \mid h_t,\, a_t)]\,.$$

The laboratory supervision pipeline pairs each target panel with the information available before its result timestamp, including prior treatments (Appendix A.1).

Privileged On-Policy Self-Distillation (OPSD). We adapt on-policy self-distillation (Zhao et al., 2026) to clinical observation prediction with a fixed privileged teacher. Let $q_t = (h_t, a_t)$ denote the student's question, containing the visible clinical history and requested action. The student samples an observation response $\hat{o}_t = (y_{t,1}, \ldots, y_{t,L_t})$ from $p_\theta(\cdot \mid q_t)$. At token position k, both models condition on the same student-generated prefix $y_{t,<k}$; the teacher additionally receives clinical evidence c_t^+. For the recorded target o_t, subsequent examination results $F_{>t}^{\mathrm{exam}}$, and final diagnosis d, the teacher-only hints are

$$c_t^+ = \begin{cases} o_t, & \text{Base OPSD,} \\ (o_t, F_{>t}^{\mathrm{exam}}), & \text{Future Hint,} \\ (o_t, F_{>t}^{\mathrm{exam}}, d), & \text{Future+Dx Hint.} \end{cases}$$

Here, future includes all later examination results in the same encounter, and Dx denotes its final diagnosis. The aligned next-token distributions are

$$p^S_{t,k} = p_\theta(\cdot \mid q_t, y_{t,<k}),$$
$$p^T_{t,k} = p_{\bar{\theta}}(\cdot \mid q_t, c_t^+, y_{t,<k}),$$

where $\bar{\theta}$ stays fixed throughout distillation. The teacher computes a next-token distribution at each student prefix, providing supervision along the student's own response. We update only the student using token-level Jensen–Shannon divergence (JSD):

$$\mathcal{L}_{\mathrm{OPSD}} = \mathbb{E}_{\hat{o}_t \sim p_\theta(\cdot \mid q_t)}\left[\frac{1}{L_t}\sum_{k=1}^{L_t} \mathrm{JSD}_\beta\big(\mathrm{sg}[p^T_{t,k}], p^S_{t,k}\big)\right].$$

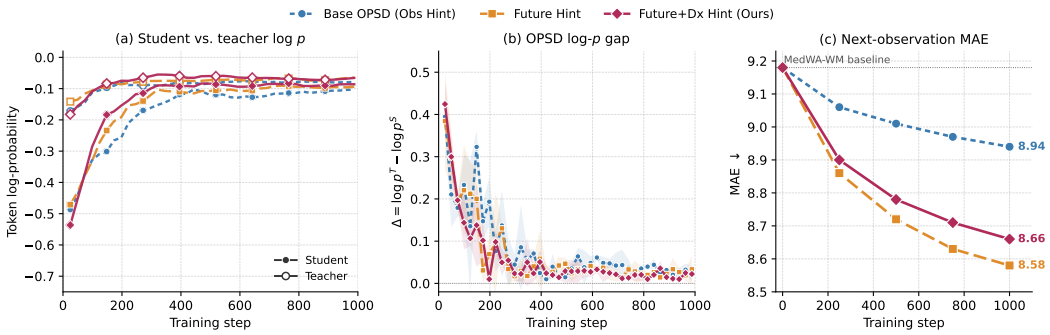

Figure 4: Effect of the teacher-only hint c_t^+ under OPSD. **(a)** Student vs. teacher token log-probability (filled markers: student; hollow: teacher); **(b)** student–teacher gap $\Delta = \log p^T - \log p^S$; **(c)** held-out next-observation MAE. Base OPSD gives the teacher o_t; Future Hint adds all later examination results; Future+Dx Hint also adds the final diagnosis. Both future-hint variants reduce held-out MAE relative to Base OPSD.

Here, L_t counts response tokens, sg stops teacher gradients, and β weights the JSD mixture (Appendix C). On-policy refers to student-generated response prefixes. Privileged evidence is restricted to the teacher during training; inference uses only q_t and the student's own prefix.

Integrating World Model into Replay. The world model conditions on the current interaction history, including observations already revealed through replay. When replay contains the requested feedback, the environment returns the real observation o_t^{real} directly. When the selected action is valid under the action schema but lacks recorded feedback, the world model fills in $\hat{o}_t$ from the routing rule in Section 4.1:

$$\hat{o}_t \sim p_\theta(\cdot \mid h_t,\, a_t).$$

This routing preserves recorded evidence wherever available. OPSD trains on student-generated response prefixes, while subsequent recorded feedback supplies grounding during multi-turn interaction. These mechanisms do not impose a bound on cumulative prediction error. Routing depends on feedback availability. The world model returns structured observations without a calibrated clinical confidence score, and token probabilities are not used as correctness estimates.

4.3 MedWA-Agent: Learning for Multi-Turn Clinical Planning

SFT and RL Training. Behavior cloning alone does not teach the agent how to compare alternatives beyond the observed clinician action. We therefore train **MedWA-Agent** on top of MedWA-Gym and MedWA-WM for multi-turn clinical planning. The agent is first initialized with SFT on patient trajectories and then optimized with reinforcement learning in the MedWA-Gym hybrid environment using group relative policy optimization (GRPO). At each turn t, the policy π_ϕ selects an action $a_t \in \mathcal{A}$ from the currently available history h_t. Here, ϕ denotes agent parameters, separate from the world model; the action space is defined in Section 4.1. An episode terminates on final diagnosis, an explicit stop, or the turn limit.

Environment-aligned reward. The episode-level reward decomposes into bundle matching, diagnosis correctness, and stepwise adjustments:

$$R = \sum_{g \in \mathcal{G}} \mathrm{F1}\big(B_g^{\text{pred}}, B_g^{\text{gold}}\big) \;+\; \mathbb{I}[\text{dx_match}] \;+\; \sum_{t=1}^{T} \Delta_t,$$

where $T \le T_{\max}$ is the realized episode length, F1 is the harmonic mean of bundle precision and recall, and $\mathcal{G} = \{\text{lab}, \text{imaging}, \text{medication}, \text{treatment}\}$, B_g^{pred} is the set of requested bundles in family g, and B_g^{gold} is the corresponding requestable set in the case record. The indicator $\mathbb{I}[\text{dx_match}]$ gives a diagnosis bonus when the final diagnosis matches the reference label, while Δ_t collects stepwise adjustments including penalties for invalid actions, unmatched bundle requests, duplicate

requests, and additional request steps, as well as bonuses for a valid diagnosis attempt and a correct final diagnosis. If the episode ends without a diagnosis, an additional missing-diagnosis penalty is applied.

Learning in the MedWA-Gym hybrid environment. At each turn, the policy conditions on the updated history containing replayed and generated observations. Later recorded evidence therefore allows it to reconsider decisions informed by an earlier generated result, although recovery from an erroneous observation is not guaranteed. The terminal diagnosis reward uses the recorded label, keeping the learning target anchored to the case reference even when intermediate feedback is generated. This design supports sequential evidence gathering under imperfect feedback; its aggregate effect is evaluated in Section 5.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Training and rollout. Both MedWA-WM and MedWA-Agent use Qwen3-4B (Yang et al., 2025). Agent training uses SFT and then GRPO (Shao et al., 2024) in the MedWA-Gym hybrid environment, with sampled cases from MedWA-118K in both stages. SFT uses LLaMA-Factory 0.8 and RL uses veRL 0.9, both on 4×A40 GPUs. The reward combines bundle-family matching, diagnosis correctness, and penalties for invalid or redundant actions. During rollout, the environment returns recorded observations when available and uses MedWA-WM for valid requests without recorded feedback.

OPSD protocol. MedWA-WM undergoes SFT followed by distillation with a fixed teacher. Each student response is appended to the student question and to the teacher's hint-augmented question, aligning their next-token distributions for JSD. Base, Future, and Future+Dx specify the teacher evidence above; only the student receives updates. Appendix C details the protocol.

Baselines and metrics. We evaluate general-purpose LLMs (GPT-5.1, Gemini 3 Pro, Kimi K2, DeepSeek-V3.1, Qwen3 variants, and GLM-4.7) and medical LLMs (MedGemma-1.5-4B, MedGemma-27B, and Baichuan-2-7B). Each model predicts next observations in the world-model comparison and uses the same prompt and action interface in the multi-turn comparison, without task-specific fine-tuning. World-model scores evaluate numeric laboratory predictions using thresholded success rates (S@10/15/20/25), mean absolute error (MAE), mean absolute percentage error (MAPE), and coverage. Agent metrics are average turns, bundle-level F1 for laboratory, imaging, treatment, and medication requests, and final diagnosis accuracy. Appendix D defines the metrics and aggregation rules.

5.2 WORLD MODEL EVALUATION

Prediction performance and OPSD ablation. Table 1 shows that all MedWA-WM variants outperform the prompted baselines across the reported next-observation metrics. The unaugmented MedWA-WM row is the SFT model; subsequent rows add OPSD with the indicated hints. Our default Future+Dx model reaches 80.14 on S@25 and 88.36% coverage, compared with 68.74 and 79.18% for Gemini 3 Pro; its MAE is 8.66 versus 13.86. Base OPSD improves the unaugmented MedWA-WM on every metric. Among the hint variants, Future Hint has the lowest MAE (8.58) and MAPE (17.92), whereas Future+Dx has the highest S@20 (75.78), S@25 (80.14), and coverage (88.36%). Category-level and signed-error analyses appear in Appendix E.1 (Figure 5).

Laboratory case analysis. Table 2 shows three laboratory chemistry panels. In Case A, predicted values closely match the recorded panel. In Case B, the model predicts sodium at 113 versus the recorded 114 and tracks the low chloride and BUN values. Case C illustrates a limitation despite full component coverage: glucose is 430 versus 442, but potassium is 4.0 versus 5.0 and anion gap is 12 versus 10. These errors reduce the case's within-10% score to 82% and could affect subsequent decisions. Coverage alone therefore does not establish that a generated panel is clinically reliable.

Table 1: World-model performance on held-out next-observation prediction. Best results are in bold and second-best results are underlined.

Model	S@10 ↑	S@15 ↑	S@20 ↑	S@25 ↑	MAE ↓	MAPE ↓	Coverage ↑
General LLMs							
GPT-5.1 (Singh et al., 2025)	43.06	52.99	61.10	66.93	15.00	57.89	76.03
Gemini 3 Pro (Google DeepMind, 2025)	45.63	55.21	63.08	68.74	13.86	50.72	79.18
Kimi K2 (Team et al., 2025)	39.71	49.36	57.44	63.15	18.42	65.37	75.46
Qwen3-235B-A22B (Yang et al., 2025)	34.31	45.10	52.31	58.12	131.02	224.92	73.80
DeepSeek-V3.1 (Liu et al., 2024)	33.33	43.25	50.50	55.24	554.98	392.71	69.92
Qwen3-30B-A3B (Yang et al., 2025)	32.77	42.60	50.66	55.72	509.95	836.51	73.32
GLM-4.7 (Glm et al., 2024)	31.56	40.30	46.99	51.08	492.67	187.64	55.20
Medical LLMs							
MedGemma-1.5-4B (Sellergren et al., 2025)	31.82	39.76	43.61	45.09	371.77	109.34	70.94
MedGemma-27B (Sellergren et al., 2025)	45.23	55.87	63.54	68.96	13.87	49.82	79.53
Baichuan-2-7B (Yang et al., 2023)	29.47	35.30	41.29	47.05	24.87	80.95	70.26
MedWA-WM	59.44	68.43	73.96	78.31	9.18	19.50	86.00
+ Base OPSD	60.02	69.01	74.52	78.99	8.94	18.88	86.71
+ Future Hint	**61.21**	**70.16**	<u>75.47</u>	<u>79.73</u>	**8.58**	**17.92**	<u>87.52</u>
MedWA-WM + Future+Dx OPSD (Ours)	<u>60.96</u>	<u>69.94</u>	**75.78**	**80.14**	<u>8.66</u>	<u>18.10</u>	**88.36**

Table 2: Qualitative next-observation predictions on three laboratory chemistry panels, with chief complaint (*CC*) and final diagnosis (*Dx*). Pred and Gold denote predicted and recorded values; cells with $> 10\%$ error are highlighted in red. "–" denotes a component that is not reported in that panel.

	Case A (mild)		**Case B** (severe, stable)		**Case C** (few outliers)	
Component	Pred	Gold	Pred	Gold	Pred	Gold
Anion Gap	17	18	11	11	12	10
BUN	12	11	3	3	39	37
Calcium	7.8	7.7	8.4	7.9	8.7	8.6
Chloride	107	107	80	80	96	97
CO_2	21	21	23	23	24	25
Creatinine	1.25	1.21	0.67	0.66	0.72	0.76
eGFR (AA)	–	–	125	130	100	102
eGFR (2021)	70	70	108	113	87	89
Glucose	185	185	100	102	430	442
Potassium	3.1	3.1	3.7	3.5	4.0	5.0
Sodium	145	146	113	114	132	132

A (CSN_99826386): *CC* fatigue, dry mouth; *Dx* dehydration with mild hyperglycemia. *cov*=1.00, *MAE*=0.38, *MAPE*=1.9%, *w10*=100%.
B (CSN_99002776): *CC* nausea, confusion, weakness; *Dx* severe hyponatremia (SIADH suspected). *cov*=1.00, *MAE*=1.25, *MAPE*=2.2%, *w10*=100%.
C (CSN_99135310): *CC* polyuria, polydipsia, altered mental status; *Dx* uncontrolled type-2 DM. *cov*=1.00, *MAE*=2.10, *MAPE*=5.7%, *w10*=82%.

5.3 Out-of-Distribution World-Model Evaluation

Transfer to MIMIC-IV-ED. We evaluate MedWA-WM on 120 MIMIC-IV-ED emergency cases spanning 12 disease categories, converted to the MedWA-Bench interaction schema (Appendix A.2). Under the same seven metrics, MedWA-WM leads the three prompted baselines in Table 3: S@10 is 52.68 versus 42.83 for Gemini 3 Pro, MAE is 11.56 versus 15.31, and coverage is 82.14% versus 77.12%. Performance is lower than on MedWA-Bench, but the model retains an advantage on this unseen corpus without corpus-specific tuning.

5.4 Multi-Turn Clinical Planning

Agent performance and world-model contribution. Table 4 shows that MedWA-Agent has the highest bundle-level F1 in all four action families and the highest diagnosis accuracy (44.92%). Compared with the same SFT+GRPO pipeline without MedWA-WM, diagnosis accuracy rises from 42.73% to 44.92%, laboratory F1 from 78.42 to 82.28, and treatment F1 from 66.21 to 69.55.

Table 3: Out-of-distribution world-model evaluation on MIMIC-IV-ED. Next-observation predictions are compared under the same schema and metrics as MedWA-Bench (Table 1). Best in bold.

Model	S@10 ↑	S@15 ↑	S@20 ↑	S@25 ↑	MAE ↓	MAPE ↓	Cov. ↑
GPT-5.1 (Singh et al., 2025)	40.62	50.38	58.47	64.71	16.48	62.35	74.52
Gemini 3 Pro (Google DeepMind, 2025)	42.83	52.41	60.36	66.27	15.31	56.84	77.12
Kimi K2 (Team et al., 2025)	37.91	46.72	54.18	60.53	20.26	71.63	73.68
Ours (MedWA-WM)	**52.68**	**61.94**	**68.47**	**73.62**	**11.56**	**28.74**	**82.14**

Table 4: Performance on MedWA-Bench. F1 and diagnosis accuracy are reported as percentages; best scores are bold. Turns measure interaction length.

Model	Avg. Turns	Lab ↑	Img ↑	Treat. ↑	Med. ↑	Diag. Acc. ↑
General LLMs						
GPT-5.1 (Singh et al., 2025)	2.39	42.27	34.08	7.40	6.65	39.09
Gemini 3 Pro (Google DeepMind, 2025)	4.35	57.83	55.64	17.92	16.24	41.36
Kimi K2 (Team et al., 2025)	4.78	55.14	51.72	20.41	18.63	39.55
DeepSeek-V3.1 (Liu et al., 2024)	5.57	63.77	59.03	6.85	6.52	37.73
Qwen3-30B-A3B (Yang et al., 2025)	5.05	51.73	63.76	2.90	4.53	34.55
Qwen3-235B-A22B (Yang et al., 2025)	3.92	64.64	73.15	24.38	22.26	40.00
GLM-4.7 (Glm et al., 2024)	3.81	55.58	59.37	8.16	10.60	31.42
Medical LLMs						
MedGemma-1.5-4B (Sellergren et al., 2025)	14.47	0.18	0.23	0.27	0.00	0.00
MedGemma-27B (Sellergren et al., 2025)	5.63	58.47	54.96	17.53	14.82	41.81
Baichuan-2-7B (Yang et al., 2023)	12.16	10.83	6.90	1.33	2.48	10.26
SFT + GRPO w/o WM	10.17	78.42	75.68	66.21	45.37	42.73
MedWA-Agent (Ours)	8.80	**82.28**	**76.92**	**69.55**	**48.15**	**44.92**

Average interaction length falls from 10.17 to 8.80 turns, close to the 8.73-turn mean of the recorded benchmark trajectories. Category and diagnosis-timing analyses appear in Appendix E.2.

Need for world-model completion. Across 220 rollout cases, 70 (31.8%) contain at least one laboratory or imaging request without recorded feedback, and 16 (7.3%) contain two or more. The 90 missing-feedback requests comprise 64 laboratory and 26 imaging requests, averaging 0.41 per case (Appendix F, Figure 9). Thus, nearly one-third of cases encounter a feedback gap, motivating world-model completion of requested observations beyond recorded trajectories. Together with the WM ablation, these results support the utility of learned feedback for multi-turn agent training.

Reliability and validation scope. The WM ablation measures the net effect of incorporating learned feedback; the performance cost of an individual WM error remains unquantified. Quality control includes source-record filtering and chronological context construction (Appendix A.1), complemented by held-out observation evaluation on two corpora. These checks assess data integrity and predictive fidelity. Generated branches have not undergone independent clinician validation; their role here is to support agent-learning research. Appendix F clarifies the evaluation scope.

6 Conclusion

We presented MedAgentWorld, combining EHR replay, a world model trained with privileged on-policy self-distillation, and reinforcement learning for multi-turn clinical planning. MedWA-WM leads the compared baselines on next-observation prediction in both MedWA-Bench and MIMIC-IV-ED. MedWA-Agent achieves the highest diagnosis accuracy and bundle-level F1 among the evaluated agents. These results support hybrid replay and learned feedback as a practical environment for studying sequential clinical agent learning.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Jinwei Gu, Siddharth Gururani, Haoxiang Wang, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

David Bani-Harouni, Chantal Pellegrini, Ege "Ozsoy, Matthias Keicher, and Nassir Navab. Language agents for hypothesis-driven clinical decision making with reinforcement learning. *arXiv preprint arXiv:2506.13474*, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801, 2025.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Sherjil Ozair, Jingwei Zhang, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, and Scott Reed. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Google DeepMind. Gemini 3, 2025. https://deepmind.google/technologies/gemini/.

David Ha and J"urgen 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.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi: 10.1056/AIdbp2500144.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. Technical report, PhysioNet, 2023.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A Kim. Mc-med, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12(1):1094, 2025.

Matthieu Komorowski, Leo Anthony Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, 642(8067):451–457, 2025. doi: 10.1038/s41586-025-08869-4.

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

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.

Mila Nambiar, Supriyo Ghosh, Priscilla Ong, Yu En Chan, Yong Mong Bee, and Pavitra Krishnaswamy. Deep offline reinforcement learning for real-world treatment optimization applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4673–4684, 2023.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P. Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind Blog, 2025. Accessed 2026-05-03.

Aamer Abdul Rahman, Pranav Agarwal, Vincent Michalski, Rita Noumeir, and Samira Kahou. Empowering clinicians with MeDT: A framework for sepsis treatment. In *NeurIPS 2023 Workshop on Goal-Conditioned Reinforcement Learning*, 2023.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020. doi: 10.1038/s41586-020-03051-4.

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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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. doi: 10.1038/s41586-023-06291-2.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. doi: 10.1038/s41591-024-03423-7.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621. Association for Computational Linguistics, 2024.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Rui Tu, Zhpeng Luo, Chuanliang Pan, Zhong Wang, Jie Su, Yu Zhang, and Yifan Wang. Offline safe reinforcement learning for sepsis treatment: Tackling variable-length episodes with sparse rewards. *Human-Centric Intelligent Systems*, 5:63–76, 2025a. doi: 10.1007/s44230-025-00093-7.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Hyatt Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle K. Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 2025b. doi: 10.1038/s41586-025-08866-7.

Yuan Wang, Anqi Liu, Jucheng Yang, Lin Wang, Ning Xiong, Yisong Cheng, and Qin Wu. Clinical knowledge-guided deep reinforcement learning for sepsis antibiotic dosing recommendations. *Artificial Intelligence in Medicine*, 150:102811, 2024. doi: 10.1016/j.artmed.2024.102811.

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, 2026.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

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*, 2025.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys*, 55(1):1–36, 2021. doi: 10.1145/3477600.

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*, 2026.

APPENDIX

Guide to the appendix. Appendix A covers data construction; Appendix B describes the interaction interface and replay prototype; Appendix C details world-model training; Appendix D defines evaluation metrics; Appendix E presents supplementary results and cases; and Appendix F discusses feedback coverage and reliability.

A DATA AND TRAJECTORY CONSTRUCTION

A.1 MC-MED SUPERVISION PIPELINE

MC-MED records are converted into two supervision streams: laboratory-value question answering (QA) for MedWA-WM and multi-turn dialogues for MedWA-Agent. Table 5 summarizes the shared case-construction stages and the two branches.

Table 5: End-to-end data curation pipeline. Stages 3A/4A produce laboratory-value QA and its SFT rendering (used to train and evaluate MedWA-WM). Stages 3B/4B/5B produce the interactive episode, its multi-turn dialogue supervision, and the per-turn expansion (used to train MedWA-Agent).

Stage	Input	Main operation	Output
1	Raw clinical CSVs	Aggregate multi-table rows per encounter	Case record
2	Case record / manifest	Quality filter, stratified sample, train/test split	Case dataset
3A	Case dataset	Extract target lab panels, build (context, panel) QA	Lab-value QA
4A	Lab-value QA	Render into LLM-ready instruction/response format	Lab SFT
3B	Case dataset	Convert to interactive episode structure	Interactive episode
4B	Interactive episode + order timeline	Generate multi-turn dialogue supervision	Multi-turn SFT
5B	Multi-turn SFT	Split each dialogue into per-turn training samples	Turn-level SFT

Case construction and selection (Stages 1–2). Records from visits, orders, imaging, laboratories, medical history, medications, and waveform summaries are linked by encounter identifier (CSN) and sorted by clinical time. Each case stores presentation context, a timeline, bundle identifiers, observations, and terminal labels. Selection uses either quality-score ranking with disease stratification or a predefined split manifest. Split, category, rank, score, and bucket metadata are retained.

World-model supervision (Stages 3A–4A). Each retained laboratory panel becomes one QA example. Its context includes only information available before the result timestamp: presentation findings, background history, prior orders, imaging, treatments, and laboratory results. Panels with insufficient numeric components are filtered out. The requested panel forms the question, and its recorded component values form the supervised response. Examples are rendered as instruction/response pairs for SFT.

Agent supervision (Stages 3B–5B). Each case supplies an initial context, requestable bundles, a feedback map and timeline, and terminal diagnostic labels. The recorded order sequence is rendered as a dialogue: request one or more bundles, receive available feedback, and continue until diagnosis. Each assistant turn then becomes a training example with the preceding dialogue as context.

Reproducibility. The source release, split manifest, panel list, minimum component count, bundle taxonomy, and disease-category map determine the pipeline outputs. Both supervision branches use the selected case dataset.

A.2 OUT-OF-DISTRIBUTION EVALUATION: MIMIC-IV-ED

The MIMIC-IV-ED evaluation (Table 3) uses 120 emergency cases, with 10 cases in each of 12 disease categories: abdominal/digestive, cardiovascular, respiratory, neurological, infection/fever, trauma/musculoskeletal, metabolic/laboratory, renal/urinary, vascular/limb, obstetric/gynecological, eye/ENT/oral, and other. Cases include presentation findings, history, longitudinal observations, tests,

medications, and terminal labels. We convert them into the same chronological request/feedback schema and laboratory taxonomy as MedWA-Bench, evaluating transfer without corpus-specific retraining.

B Multi-Turn Interaction and Action Encoding

The bundle-based interaction layer (`interact.py`) handles LLM requests against case-level feedback records. The event-replay prototype (`env.py`) separately provides a coarser, integer-coded interface. We describe these interfaces separately because their action encodings and transition rules differ.

B.1 Bundle-Based Interaction

Request encoding and availability. Requests use namespaced identifiers of the form `family::name`, with prefixes `lab::`, `imaging::`, `treatment::`, and `diagnostic::`. A JSON request can identify a bundle through `action` or `bundle_id`; `bundle_ids` and `treatment_bundles` support multiple requests in one turn. The parser removes repeated identifiers within a call. Control actions are `wait`, `make_diagnosis`, and `stop`. The available bundle set comes from each case's `action_space.requestable_bundles`; matching requires exact identifier equality and excludes previously requested bundles. Thus the runtime action set is case-dependent, rather than a fixed set of 20 choices.

Validation and recorded feedback. Validity depends on the calling protocol and bundle availability, without a separate clinical-appropriateness classifier. Strict JSON parsing is configurable; the implementation also supports permissive extraction. For an accepted laboratory or imaging bundle, the interaction reads `feedback_by_bundle[bundle_id].returns` and appends available results to the feedback history. If no result is present, it records the missing feedback and returns order metadata with a no-structured-feedback notice. This feedback-availability check supplies the replay branch of the hybrid routing in Section 4.1.

World-model service. MedWA-WM is trained with the same interaction interface and action format used during rollout. During agent RL, the trained world model is served through vLLM. For a valid request without recorded feedback, the rollout system sends the visible history and formatted action to this service and returns its generated observation to the agent. The interaction module handles case records and request matching, while the separately served world model supplies missing feedback; subsequent turns include the returned observation in their history.

Rewards and episode control. The default bundle-based reward sums set F1 across laboratory, imaging, medication, and treatment groups, plus a diagnosis-match indicator. Configurable adjustments handle format, requests, and diagnosis attempts. Each response increments the turn counter; `stop` terminates the episode and the configured step cap truncates it. Termination on diagnosis is controlled by a separate flag. The module defaults to a 12-step cap, permissive parsing, and no automatic termination on `make_diagnosis`; these are implementation defaults, not a specification of the experimental run configuration.

B.2 Reference Action Vocabulary and Encoding

The event-replay schema contains 20 high-level action labels (Table 6). The ordered `ACTIONS` tuple defines their integer encoding: `ACTION_TO_INDEX` maps labels to IDs 0–19, and `INDEX_TO_ACTION` provides the inverse. For example, `wait`, `order_lab_basic`, `make_diagnosis`, and `stop` have IDs 0, 1, 18, and 19. Bundle identifiers use a `family::name` format. The explicit mapping contains 51 entries: 3 diagnostic, 7 imaging, 14 laboratory, and 27 treatment identifiers. These entries map multiple bundles to the same action label; they are not 51 distinct high-level actions.

Table 6: The 20 high-level action labels, in ID order (0–19). Request actions aggregate related bundles; control and terminal actions govern episode progression.

Action name	Family	Aggregated bundles (examples)
wait	control	advance one event without ordering
order_lab_basic	lab	CBC, BMP/CMP, pregnancy, tox, endocrine, nutrition
order_lab_infection	lab	cultures, blood gas/lactate, urinalysis, COVID/flu/RSV
order_lab_cardiac	lab	troponin, BNP/NT-proBNP, 12-lead ECG, stress ECG
order_lab_coag	lab	PT/INR, PTT, D-dimer, coagulation panels
order_lab_liver_renal	lab	chem/electrolyte/renal, liver/pancreas (ALT, AST, BUN, Cr)
order_imaging_chest_xray	imaging	chest X-ray
order_imaging_head_ct	imaging	head CT (incl. stroke CTA/CTP)
order_imaging_brain_mr	imaging	head MRI / MRA
order_imaging_abdomen_us	imaging	abdominal / pelvic ultrasound
order_imaging_abdomen_ct	imaging	abdomen / pelvis CT
order_monitoring_support	care	vitals, IV line, respiratory support, mobility/safety
order_iv_fluids	care	IV fluids and electrolyte repletion
order_symptom_control	care	analgesia/sedation, antiemetics, allergy relief
order_antiinfective	care	anti-infective medications
order_consult	workflow	cardiology, neurology, surgery, OB-Gyn, ICU consult
order_admit_ward	workflow	inpatient / observation disposition
order_admit_icu	workflow	critical-care / ICU admission
make_diagnosis	terminal	commit to a final diagnostic label
stop	terminal	terminate the episode

Event mapping. The event-to-action helper first requires an event type beginning with `order_`. It reads `detail.bundle`, falling back to `detail.mapped_bundle`, and checks the explicit dictionary before examining the procedure name. For example, `lab::cbc_hematology` maps to `order_lab_basic`, whereas `lab::chem_electrolyte_renal` maps to `order_lab_liver_renal`. When the dictionary has no match, event type and uppercase procedure-name keywords determine the label: consultation maps to `order_consult`; admission maps to ICU or ward based on ICU/intensive keywords; ECG maps to the cardiac action. Imaging uses modality/anatomy matches. Laboratory matching follows infection, cardiac, coagulation, liver/renal, then basic keywords, with a basic-laboratory fallback for unmatched laboratory orders. Unsupported events return no oracle action. This helper constructs labels for recorded events; it is distinct from runtime request validation.

B.3 Event-Replay Prototype

Step and termination semantics. `EmergencyEnv` exposes `reset`/`step` and a discrete space of 20 actions. The prototype accepts integer IDs 0–19; non-integer or out-of-range inputs raise `ValueError`. Action-name lookup also rejects unknown labels. The returned `valid_action_mask` marks diagnosis unavailable after its first declaration, but `step` does not enforce this mask: a repeated diagnosis receives a penalty. A non-stop action consumes the next recorded event and updates the state regardless of whether it matches the selected action; matching affects reward. `make_diagnosis` sets a declaration flag and does not itself end the episode. Termination occurs on `stop` or event exhaustion; a configured `max_steps` causes truncation. Its default is `None`, so the prototype imposes no fixed step cap unless supplied by the caller. These semantics describe the replay prototype, not the diagnosis-terminating bundle interface used in the main evaluation.

The loader accepts nested JSON patient-flow files and bundle-aware `state_traces.jsonl`, with optional encounter-ID split filtering. At reset, initial arrival/roomed events are revealed. Observations contain `vector`, `state_summary`, `online_state`, the full `state`, and the action mask; `info` additionally contains replay metadata and reference labels.

State representation. `state_builder.py` updates each bundle's ordered, administered, and resulted counts and timestamps; demographics and presentation context are initialized at reset. `online_state.py` groups the state into patient context, evidence, care, workflow, and process

fields. These include returned-test summaries, supportive care, disposition, and event progress. The full state and metadata are distinct from the visible-history input defined in Section 4.1.

Care-response helpers. Nine categorical traces cover pain relief, gastrointestinal symptoms, hemodynamic and respiratory support, urinary decompression, bleeding control, cardiac and neurological management, and metabolic correction. Bundle-triggered rules advance these traces through baseline, started, and progressing states. Additional heuristic scores summarize stabilization, symptom relief, and pending versus returned evidence.

Reward defaults. The prototype combines negative action cost, agreement with the consumed event's mapped action, and terminal adjustments. Default `RewardConfig` values are: action-cost scale 1.0, oracle match +0.5, mismatch -0.1, first diagnosis declaration +0.4, repeated declaration -0.2, premature stop -1.5, and completion with/without a diagnosis declaration +2.0/-2.0. Action-specific costs are defined in `ACTION_COSTS`. These are prototype defaults; its declaration bonus does not evaluate diagnostic correctness and is distinct from the bundle-F1 and diagnosis-matching reward in Section 4.3.

C World-Model Training Protocol

Backbone and stages. MedWA-WM and MedWA-Agent both use Qwen3-4B. The world model first learns recorded observations through SFT, then undergoes privileged OPSD. The training-data construction is given in Appendix A.1; the three teacher-hint variants are defined in Section 4.

Inputs and fixed teacher. We adapt the OPSD formulation of Zhao et al. (2026) to structured clinical feedback. The student question q_t contains the currently available patient history and requested observation. A separate teacher prompt adds c_t^+: the recorded target observation for Base OPSD, all subsequent examination results for Future Hint, and additionally the final diagnosis for Future+Dx. The teacher parameters remain fixed throughout distillation, and only student parameters are optimized. Hints never enter the student prompt or the deployed world-model input.

Rollout and prefix alignment. The current student samples an observation response to q_t. The same sampled response tokens are appended to both prompts. At position k, causal attention restricts both models to the student prefix $y_{t,<k}$; the teacher also sees its privileged prompt. This produces two aligned next-token distributions despite the different prompt lengths. Teacher logits are computed without gradients, and distillation is evaluated on response positions, excluding prompt and padding positions. The teacher thus supplies a distribution at each student prefix; an independently sampled teacher completion is not required for this loss. Token-level generation here is distinct from the sequence of clinical actions in a MedWA-Gym episode.

JSD objective. Let $P = \mathrm{sg}[p^T_{t,k}]$, $Q = p^S_{t,k}$, and $M_\beta = \beta P + (1-\beta)Q$. The generalized Jensen–Shannon divergence is

$$\mathrm{JSD}_\beta(P,Q) = \beta D_{\mathrm{KL}}(P\|M_\beta) + (1-\beta)D_{\mathrm{KL}}(Q\|M_\beta), \qquad 0<\beta<1.$$

D_{KL} denotes Kullback–Leibler divergence; equal weighting ($\beta = 1/2$) gives the symmetric JSD. The objective averages aligned response-token divergences, with gradients flowing only through the student. Fixed teacher weights do not imply fixed targets: teacher distributions vary with the case, hint, and current student prefix.

D EVALUATION METRICS

This section formally defines every metric reported in §5 for both the world-model evaluation (Tables 1 and 3) and the multi-turn agent evaluation (Table 4). Throughout, we use $\mathcal{C}$ to denote the components inside a target laboratory panel (e.g., *Sodium*, *Potassium*), $\mathcal{P}$ to denote the set of held-out (case, panel) pairs at test time, and $|\cdot|$ to denote set cardinality.

D.1 WORLD-MODEL METRICS

For each held-out (case, panel) pair $(c, p) \in \mathcal{P}$, the world model returns a predicted panel $\hat{\mathbf{y}}^{(c,p)}$ whose keys are a subset of the ground-truth components. Let $\mathcal{C}^{(c,p)}$ be the ground-truth component set with numeric reference values $y_k^{(c,p)}$ for $k \in \mathcal{C}^{(c,p)}$, and let $\widehat{\mathcal{C}}^{(c,p)} \subseteq \mathcal{C}^{(c,p)}$ be the components actually predicted by the model. We use $\hat{y}_k^{(c,p)}$ for the prediction and evaluate prediction-error metrics on *matched* components $k \in \widehat{\mathcal{C}}^{(c,p)}$.

Coverage. Coverage measures how often the model returns a usable prediction for the components it is asked about. On a single QA it is the fraction of gold components that appear in the model output, and we macro-average over the test set:

$$\text{Coverage} \;=\; \frac{1}{|\mathcal{P}|} \sum_{(c,p)\in\mathcal{P}} \frac{|\widehat{\mathcal{C}}^{(c,p)}|}{|\mathcal{C}^{(c,p)}|}.$$

Coverage is reported as a percentage in Tables 1 and 3.

Relative error and success rate S@τ. For a matched component k, the absolute relative error is

$$e_k^{(c,p)} \;=\; \frac{|\hat{y}_k^{(c,p)} - y_k^{(c,p)}|}{\max(|y_k^{(c,p)}|,\, \varepsilon)}, \qquad \varepsilon = 10^{-9}.$$

Let $\mathcal{P}_+$ contain panels with at least one matched numeric prediction. For a threshold $\tau \in \{10, 15, 20, 25\}$ percent, the success rate S@τ is the fraction of matched components with $e_k^{(c,p)} \le \tau/100$, macro-averaged first inside each panel and then across panels:

$$\text{S@}\tau \;=\; \frac{1}{|\mathcal{P}_+|} \sum_{(c,p)\in\mathcal{P}_+} \frac{1}{|\widehat{\mathcal{C}}^{(c,p)}|} \sum_{k\in\widehat{\mathcal{C}}^{(c,p)}} \mathbb{I}\Big[e_k^{(c,p)} \le \tau/100\Big].$$

Higher is better. Coverage acts as an availability gate: components missing from the prediction are not scored, but the missing mass is already reflected in Coverage.

Mean absolute error (MAE) and mean absolute percentage error (MAPE). Let $y_k^{(c,p)}$ and $\hat{y}_k^{(c,p)}$ be the reference and prediction on matched components. MAE and MAPE are:

$$\text{MAE} = \frac{1}{|\mathcal{P}_+|} \sum_{(c,p)\in\mathcal{P}_+} \frac{1}{|\widehat{\mathcal{C}}^{(c,p)}|} \sum_{k\in\widehat{\mathcal{C}}^{(c,p)}} |\hat{y}_k^{(c,p)} - y_k^{(c,p)}|,$$

$$\text{MAPE} = \frac{1}{|\mathcal{P}_+|} \sum_{(c,p)\in\mathcal{P}_+} \frac{1}{|\widehat{\mathcal{C}}^{(c,p)}|} \sum_{k\in\widehat{\mathcal{C}}^{(c,p)}} e_k^{(c,p)}.$$

Both are lower-is-better and are reported over matched components only; MAPE is expressed as a percentage. Because clinical laboratory values live on heterogeneous scales, MAPE is the more comparable indicator across panels, while MAE aggregates raw numeric residuals across the included components and therefore has no single physical unit.

D.2 MULTI-TURN AGENT METRICS

For each test case c, MedWA-Agent produces a trajectory τ_c consisting of a sequence of turns. Requests can select bundles across four families: laboratory, imaging, treatment, and medication. A separate action commits to a final diagnosis. We denote by $\mathcal{B}_c^f$ the ground-truth set of bundles in family f from the recorded case, and by $\widehat{\mathcal{B}}_c^f$ the union of bundles in that family requested at any turn of the predicted trajectory.

Bundle-family F1 (Lab F1, Imaging F1, Treatment F1, Medication F1). For each family $f \in \{\text{lab}, \text{img}, \text{tr}, \text{med}\}$ we compute set-level precision, recall, and F1 against the reference bundles, and macro-average across cases:

$$P_c^f = \frac{|\widehat{\mathcal{B}}_c^f \cap \mathcal{B}_c^f|}{\max(|\widehat{\mathcal{B}}_c^f|,\, 1)}, \quad R_c^f = \frac{|\widehat{\mathcal{B}}_c^f \cap \mathcal{B}_c^f|}{\max(|\mathcal{B}_c^f|,\, 1)},$$

$$F1_c^f = \frac{2|\widehat{\mathcal{B}}_c^f \cap \mathcal{B}_c^f|}{\max(|\widehat{\mathcal{B}}_c^f| + |\mathcal{B}_c^f|, 1)}, \qquad \text{F1}^f = \frac{1}{|\mathcal{D}|} \sum_{c \in \mathcal{D}} F1_c^f.$$

$\mathcal{D}$ denotes the evaluation set (220 cases for MedWA-Bench). As an exception to the formula, cases with $|\mathcal{B}_c^f| = |\widehat{\mathcal{B}}_c^f| = 0$ contribute an F1 of 1 (no request expected, none issued). The same set-based bundle matching is used in the episode-level reward (Section 4.3) and in evaluation.

Diagnosis accuracy (Dx Acc). Every case is associated with a fixed set of diagnostic candidates and a single ground-truth label $d_c^\star$. If the agent's final `make_diagnosis` action selects a label $\hat{d}_c$, the case contributes $\mathbb{I}[\hat{d}_c = d_c^\star]$; if the trajectory ends without a diagnosis, it contributes 0. Diagnosis accuracy is then the case-macro average:

$$\text{Dx Acc} = \frac{1}{|\mathcal{D}|} \sum_{c \in \mathcal{D}} \mathbb{I}[\hat{d}_c = d_c^\star].$$

This is exact-match on the terminal label, not partial credit on ICD-10 chapters.

Average turns. For each case c let T_c be the total number of turns in the executed trajectory (including the terminal `make_diagnosis` or `stop`). The reported "Avg. Turns" is

$$\overline{T} = \frac{1}{|\mathcal{D}|} \sum_{c \in \mathcal{D}} T_c.$$

Trajectories are capped at a fixed budget (32 turns); truncated episodes contribute $T_c = 32$. Values close to the ground-truth trajectory mean of 8.73 are preferable, and the gap in either direction is discussed in §5.

Aggregation and edge cases. All world-model metrics use a two-stage macro average—first within a (case, panel), then across (case, panel) pairs—so that panels contribute equally regardless of the number of numeric components they contain. All agent metrics use case-macro averages so that long, complex encounters do not dominate short ones. Ties in ranked comparisons are broken by the next reported metric in the same row block (S@10 first, then S@25, then MAE) and are labelled in bold/underline as annotated in the corresponding table captions. Predictions that fail to parse as valid JSON are treated as empty predictions: they contribute 0 to Coverage on the affected panel and are excluded from S@τ/MAE/MAPE for that panel; the same rule applies to malformed `make_diagnosis` outputs, which contribute 0 to Dx Acc.

E SUPPLEMENTARY RESULTS AND CASE STUDIES

E.1 WORLD-MODEL PREDICTION ERRORS

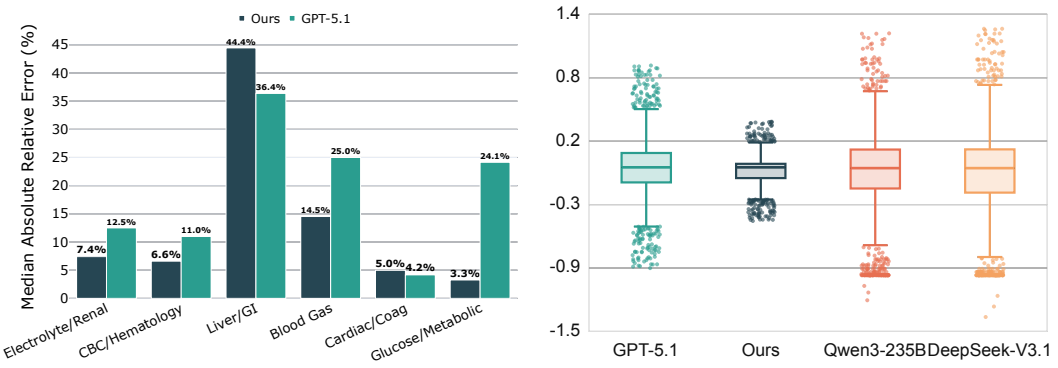

Figure 5: Left: Median absolute relative error of world-model predictions across major laboratory value categories, comparing Ours and GPT-5.1. Right: Distribution of signed relative error for world-model value prediction.

Category-Level Relative Error. Figure 5 shows that the advantage of our world model is broad rather than concentrated in a few categories. The model maintains lower relative error across most major laboratory groups, indicating that its gains come from more stable quantitative estimation instead of isolated wins on a small subset of variables.

Signed Error Distribution. Figure 5 (right) shows a tighter signed relative-error distribution for MedWA-WM than for the compared models. The boxplots summarize the direction and spread of prediction errors; they do not provide calibrated confidence for an individual generated observation (Appendix F).

Table 7: Per-case MAE broken down by disease category on MedWA-Bench. Rows are sorted by median case-level MAE. *p90* is the 90th percentile. Each case MAE aggregates component errors across its laboratory panels; columns summarize these case scores within each category.

Category	Mean	Median	p90
Eye / ENT / Dental	8.66	**6.21**	16.56
OB/GYN	7.45	6.52	13.80
Other	9.97	6.74	14.05
Renal / Urinary	8.26	7.17	12.65
Chest / Cardiac	8.17	7.47	11.55
Lab / Metabolic	10.19	7.81	20.36
Respiratory	10.30	8.05	20.40
Vascular / Extremity	9.82	8.40	17.21
Trauma / MSK	10.55	8.60	19.09
Abdominal / GI	10.68	9.21	17.97
Infection / Fever	13.96	9.39	20.23
Neurology	13.80	**9.48**	33.08

Per-category Case MAE. Table 7 reports case-level MAE by disease category. Median errors range from 6.21 for eye/ENT/dental cases to 9.48 for neurology. Infection/fever and neurology have means substantially above their medians and p90 values of 20.23 and 33.08, respectively, indicating a tail of cases with larger errors. These summaries describe heterogeneity in prediction error; they do not identify its clinical cause or establish reliability for an individual case.

E.2 Agent Diagnosis Patterns

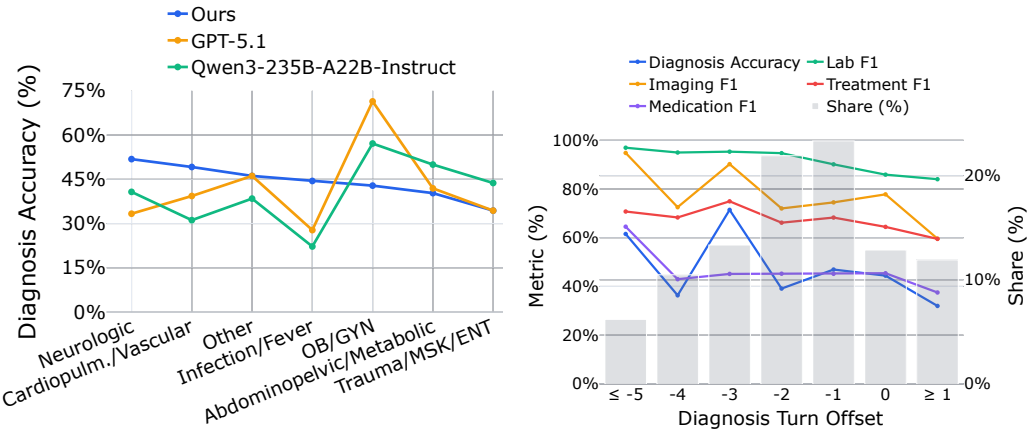

Figure 6: Left: Diagnosis accuracy across aggregated disease categories on MedWA-Bench, comparing Ours, GPT-5.1, and Qwen3-235B-A22B. Right: Model performance as a function of diagnosis turn offset, defined as ground-truth diagnosis turn minus predicted diagnosis turn. Negative values indicate that the model commits later than the ground-truth trajectory.

Category-Specific Diagnosis Accuracy. Figure 6 (left) shows that the downstream gains are not uniform across disease groups. Our method is strongest on several common emergency-care categories that require iterative evidence integration, while some abdominopelvic, metabolic, and trauma-related groups remain weaker and still favor a strong general baseline.

Diagnosis Timing Trade-off. Figure 6 (right) reports diagnosis accuracy as a function of the offset between the ground-truth commitment turn and the model's commitment turn. Performance is strongest when the agent waits slightly longer before issuing the final diagnosis, whereas earlier decisions are associated with weaker diagnosis and intervention quality. This association is consistent with the value of additional evidence, although the plot does not isolate the effect of delaying a decision.

E.3 MULTI-TURN CASE STUDIES

The following cases show how an agent request passes through recorded-feedback replay and, when a result is missing, world-model completion. Each timeline pairs the agent's requests with the feedback it receives. **Recorded** denotes feedback from the case record; **Generated** denotes MedWA-WM output, highlighted in orange. Clinical names replace most action identifiers for readability; the completed bundle retains its exact identifier. The first case illustrates missing imaging feedback, and the second illustrates missing laboratory feedback. Correct final labels describe the observed rollouts; they do not validate the generated findings or establish that completion caused the correct diagnosis.

Case 1: Missing ultrasound feedback during a heart-failure workup

Presentation. A 56-year-old with chest pain and breathing difficulty; prior thoracic aortic dissection and combined systolic/diastolic heart failure. At presentation: heart rate 70, respiratory rate 18, oxygen saturation 99%, blood pressure 129/70; triage acuity 3 (urgent).

Turn	Agent request / action	Feedback available to the agent
1	Infection screen and 12-lead ECG.	**Recorded:** No abnormal microbiology components; ECG documented.
2	Blood count and chemistry; isolation / communication bundles.	**Recorded:** Microcytic anemia (hemoglobin 9.5, hematocrit 32.5, MCV 80); chemistry CO_2 31, flagged high.
3	Cardiac biomarkers and chest X-ray.	**Recorded:** NT-pro-BNP 801, flagged elevated; chest X-ray report available.
5	Chest CT angiography for pulmonary embolism; cardiovascular/renal care and cardiology consultation.	**Recorded:** Cardiovascular/renal treatment and cardiology consultation documented.
6	Point-of-care ultrasound to assess cardiac function; inpatient disposition.	**Feedback gap:** No structured result for `imaging::pocus_limited`.
6	**WM completion**	**Generated:** Reduced left-ventricular ejection fraction and pulmonary congestion. This response is appended to the interaction history.
7	Submit final diagnosis.	**Agent:** Acute combined systolic and diastolic heart failure (label D).

Outcome. The prediction matches the recorded diagnosis: acute combined systolic and diastolic heart failure (ICD-10 I50.41). Final reward: 4.66. Bundle F1: laboratory 0.73, imaging 1.00, medication 1.00, treatment 0.93.

Key interaction. At turn 6, an accepted ultrasound request has no recorded result. WM supplies the missing response before the agent submits its diagnosis at turn 7. Only selected turns are shown; turn 4 is omitted.

Figure 7: **Imaging-feedback completion** in case CSN_99052807. The shaded row separates generated ultrasound findings from recorded observations.

Case 2: Missing cardiac biomarkers during a respiratory workup

Presentation. An 82-year-old with shortness of breath; no documented past medical history or chronic medications. At presentation: heart rate 88, respiratory rate 18, oxygen saturation 97% on room air, blood pressure 110/56; triage acuity 3 (urgent).

Turn	Agent request / action	Feedback available to the agent
1	Infection screen and 12-lead ECG.	**Recorded:** SARS-CoV-2 positive; ECG documented.
2	Blood count and chemistry; isolation and inpatient notification.	**Recorded:** Neutrophilia; sodium 132, chloride 97, and albumin 2.9 flagged low; AST 52 and ALT 76 elevated.
3	Cardiac biomarkers and blood gas / lactate.	**Recorded:** Venous blood gas: pH 7.44, HCO_3 26.1, pCO_2 38.2, pO_2 46. **Feedback gap:** No structured result for `lab::cardiac_biomarkers`.
3	**WM completion**	**Generated:** High-sensitivity troponin-I 18 ng/L; CK-MB within reference; NT-pro-BNP 612 pg/mL. These are model outputs, not measured values in the replay record.
4	Chest X-ray.	**Recorded:** Report available; no focal consolidation flagged.
5–6	IV fluids/electrolytes; chest CT angiography for pulmonary embolism; hospitalist consultation and observation disposition.	The trajectory continues after the generated panel. No additional test findings are reproduced in this summary.
7	Submit final diagnosis.	**Agent:** Hypoxia (label F).

Outcome. The prediction matches the recorded diagnosis: hypoxia (ICD-10 R09.02). Final reward: 3.92. Bundle F1: laboratory 1.00, imaging 1.00, medication 0.00, treatment 0.92.

Key interaction. Turn 3 contains both an available result (venous blood gas) and a missing panel (cardiac biomarkers). Replay returns the former and WM completes the latter; the agent then continues requesting tests before submitting its diagnosis at turn 7.

Figure 8: **Laboratory-feedback completion** in case CSN_99842446. The shaded row identifies generated cardiac biomarkers within an otherwise recorded-feedback trajectory.

Abbreviations used in cases and tables. CBC: complete blood count; CMP: comprehensive metabolic panel; Hgb/Hct: hemoglobin/hematocrit; BUN: blood urea nitrogen; eGFR: estimated glomerular filtration rate; ECG: electrocardiogram; POCUS: point-of-care ultrasound; LVEF: left-ventricular ejection fraction; PMH: past medical history; SIADH: syndrome of inappropriate antidiuretic hormone secretion; DM: diabetes mellitus. Category labels use ENT (ear, nose, and throat), OB/GYN (obstetrics and gynecology), MSK (musculoskeletal), and GI (gastrointestinal). In Table 2, cov denotes component coverage and w10 denotes the fraction of evaluated values within 10% of the recorded value.

F Feedback Coverage and Reliability

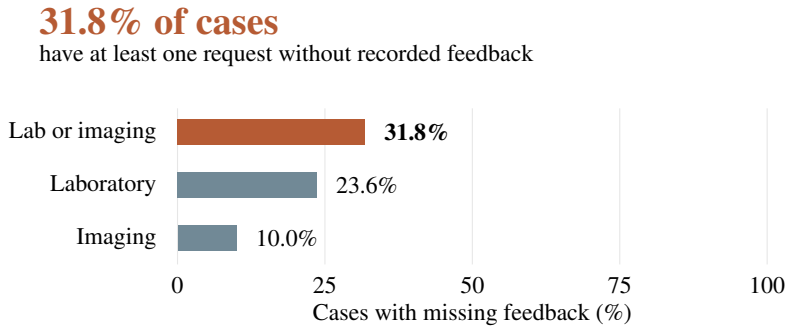

Figure 9: **Feedback gaps affect nearly one-third of rollout cases.** Bars show the fraction of all 220 cases with at least one request lacking recorded feedback. Laboratory and imaging groups overlap; their union accounts for 31.8% of cases. These are case-level rates, not step-level invocation rates.

Observed feedback gaps. Figure 9 shows that 31.8% of cases encounter missing recorded feedback. For laboratory and imaging combined, the numbers of cases with 0, 1, 2, 3, and 4 missing requests are 150, 54, 13, 2, and 1, respectively. Thus, 70/220 cases (31.8%) require at least one feedback completion, and 16/220 (7.3%) require multiple completions. The corresponding counts are 168, 42, 8, 2, and 0 for laboratory requests and 198, 19, 2, 1, and 0 for imaging requests. Summing missing requests yields 90 in total: 64 laboratory and 26 imaging, or 0.409, 0.291, and 0.118 per case, respectively. These are request counts and case-level exposure statistics. A single step may request several bundles, so they do not determine the fraction of steps invoking MedWA-WM or the number of model calls. The distribution establishes the need for feedback completion, while its downstream value is assessed by the agent ablation.

Sequential reliability. MedWA-Gym returns recorded observations whenever available and uses MedWA-WM to complete missing feedback. OPSD matches teacher and student distributions on student-generated response prefixes. Later recorded observations provide additional evidence for subsequent predictions and agent decisions. These mechanisms motivate the hybrid design; aggregate evaluation does not establish cumulative error growth or reliable recovery after an incorrect observation. The 2.19-point diagnosis gain in Table 4 measures the net benefit of incorporating WM feedback under the evaluated protocol, without isolating the effect of individual errors.

Confidence and clinical validation. The interface returns structured observations without calibrated clinical confidence estimates; token likelihood and teacher–student agreement are training signals rather than clinical correctness probabilities. The router uses feedback availability. Source-data controls include identifier and label filtering, chronological context construction, and filtering of panels with insufficient numeric components (Appendix A.1). Held-out next-observation evaluation measures predictive fidelity on recorded outcomes, including transfer to MIMIC-IV-ED. These checks support the use of MedWA as a research environment, but do not establish the clinical validity of unobserved action outcomes. Generated trajectories have not undergone independent clinician validation, and the reported results do not establish readiness for clinical deployment.

Scope of action matching. The framework is designed to align agent behavior with historically recorded clinical trajectories, so the reward signal reflects what physicians actually did rather than what they optimally should have done. Expert review would help assess decisions that depart from historical practice. Finally, actions are defined at the bundle level rather than at the level of individual orders, a deliberate design choice to keep the action space tractable; enumerating fine-grained clinical options would yield thousands of candidates, increasing the complexity of action selection. However, this means dose selection and fine-grained administration timing remain outside the present scope.