Medical World Models

Clin-JEPA, EHRWorld, and CLARITY

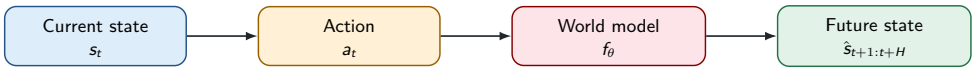
Jiaxi Liu

School of Life Sciences, Peking University

June 15, 2026

Outline

World Models: From Game Intuition to Clinical Intuition



- In games: screen + control input $\rightarrow$ the next visual frame.
- In robotics: sensors + motor command $\rightarrow$ the next pose.
- In medicine: patient state + treatment/test/medication $\rightarrow$ future condition.

Key point: a world model emphasizes dynamic transition, not only static classification.

Translating World Models into Clinical Language

Machine learning concept	Clinical meaning
State	The patient's real but only partly observed physiological state, such as organ function, tumor burden, or inflammation.
Observation	Recorded data, such as vital signs, laboratory tests, imaging, and clinical notes.
Action	Orders, medications, surgery, radiotherapy, tests, and nursing interventions.
Transition	State change caused jointly by natural disease course and treatment.
Rollout	Multi-step simulation into the future, for example the next 48 hours in the ICU.
Value / risk	Mortality, recurrence, organ failure, length of stay, or survival probability.

Why Medical World Models Are Difficult

- **The state is not fully observed**: the real physiology is only seen indirectly through EHR, imaging, and tests.
- **Actions are not randomized**: physicians choose treatment according to severity, so EHR data are strongly confounded.
- **Time is irregular**: ICU data may be hourly, while oncology follow-up may be weeks or months apart.
- **Errors accumulate**: one-step prediction may look good, but multi-step simulation can drift.

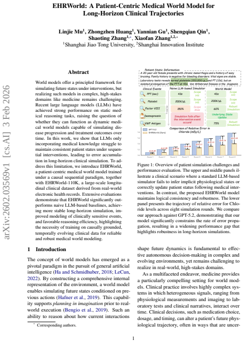
What is EHR?

EHR = Electronic Health Record. It is time-series data left in hospital information systems: diagnoses, vital signs, laboratory tests, medication orders, procedures, nursing records, and progress notes.

Safety boundary

EHR is a recorded observation, not the patient's true physiological state itself. These papers show predictive and simulation ability on research data; this is not the same as a deployable clinical system.

Putting the Three Papers on One Map



Clin-JEPA
ICU EHR latent trajectory prediction

EHRWorld
Patient-centered full-stay EHR event simulation

CLARITY
Latent-space trajectory under oncology treatment

Outline

What Problem Does Clin-JEPA Address?

Core question:
Can we train an EHR backbone that supports both future patient-state simulation and multiple ICU risk tasks, instead of doing separate feature engineering and fine-tuning for each task?

- Input: hourly clinical observations, interventions, and admission demographics from MIMIC-IV ICU.
- Representation: each hourly EHR text is encoded into a 4096-dimensional latent embedding.
- World model: a latent trajectory predictor is retained for autoregressive rollout in latent space.
- Challenge: joint training of the encoder and predictor can easily collapse or drift.

JEPA Intuition: Predict Representations, Not Raw Data

- JEPA = Joint-Embedding Predictive Architecture.
- It is a self-supervised framework that **predicts in representation space**: inputs are encoded into latent embeddings, then the model predicts embeddings of masked or future states.
- It does not directly reconstruct pixels, waveforms, or raw text, so it focuses more on high-level semantics and dynamics.
- Clin-JEPA transfers this idea by replacing image patches with hourly ICU EHR and the target with the next-hour patient-state representation.

Three components

1. **context encoder**: encodes the known part

$$z_x = E_{\text{context}}(x_{\text{visible}})$$

2. **target encoder**: encodes the target part, often serving as a stop-gradient or EMA teacher signal

$$z_y = E_{\text{target}}(x_{\text{target}})$$

3. **predictor**: predicts the target embedding from the context embedding

$$\hat{z}_y = P_\theta(z_x), \qquad \mathcal{L} = d(\hat{z}_y, \text{sg}(z_y))$$

From Image JEPA to EHR JEPA

Question	Image/video JEPA	Clin-JEPA
Context	Image patches, video clips	Past several hours of ICU EHR
Target	Embedding of masked patches or future frames	Next-hour patient-state embedding
Action	Usually weak or post-hoc in video	Medication, ventilation, procedures, and action text
Is the predictor kept?	Often discarded	Kept for future rollout
Main risk	Representation collapse	Collapse + online/target drift

What Does Clin-JEPA's EHR Input Look Like?

Hourly state text s_t

```
t=12h | Heart rate: 88 bpm.
MAP: 62 mmHg.
Lactate: 2.1 mmol/L.
SOFA: ...
```

Hourly action text a_t

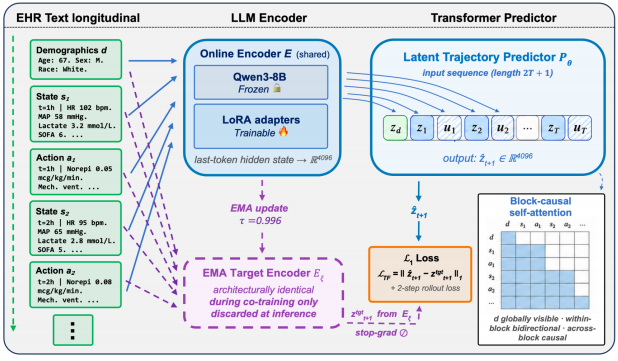
```
t=12h | Norepinephrine:
0.08 mcg/kg/min.
Propofol: 40 mcg/kg/min.
```

$$\mathcal{S} = \{d, (s_1, a_1), \ldots, (s_T, a_T)\}, \qquad T \leq 72$$

d: static demographics such as age, sex, and race;
s_t: observations at hour t, including vital signs, labs, and ventilator settings;
a_t: interventions at that hour, such as medications, procedures, and ventilation.

- Convert vital signs, laboratory results, medications, and ventilator settings into structured natural-language text.
- This reduces complex preprocessing such as missing-value imputation and normalization.
- The goal is for the LLM encoder to read hourly EHR first, then let the predictor learn cross-hour dynamics.

Clin-JEPA Architecture Overview



Five-Phase Co-Training Curriculum

# Phase	Encoder LoRA
1 Warmup	frozen
2 Co-training	trainable ⋆
3 Alignment	frozen
4 Hard sync	instant
5 Finalize	frozen

⋆ *Phase 2 alone updates the online encoder LoRA, producing dynamically grounded representations under a shared JEPA prediction objective; surrounding phases prevent representation collapse and online/target drift. Full schedule in Tab. 1; five-paradigm ablation in §5.2.*

1. **Encoder** E: based on Qwen3-8B, using LoRA to map text segments of s_t, a_t, and d into 4096-dimensional latent embeddings, denoted z_t, u_t, and z_d.
2. **Latent trajectory predictor** P_θ: an approximately 92M-parameter Transformer that takes embedding sequences in latent space and predicts the next state embedding $\hat{z}_{t+1}$.

Key Formula 1: Encoding and Rollout

$$z_t = E(s_t), \qquad u_t = E(a_t), \qquad z_d = E(d), \quad z_t, u_t, z_d \in \mathbb{R}^{4096}$$

$$\hat{z}_{t+h} = P_\theta(\tilde{z}_{1:t+h-1}, u_{1:t+h-1}, z_d), \qquad h = 1, \ldots, H$$

$$\tilde{z}_j = \begin{cases} z_j, & j \leq t \\ \hat{z}_j, & j > t \end{cases}$$

- z_t is the latent vector for the patient's state in this hour.
- u_t is the latent vector for treatment and intervention in this hour.
- During inference, the predictor can autoregressively simulate trajectories over H hours.
- $\tilde{z}_j$ uses real embeddings for historical steps ($j \leq t$) and the model's own prior predictions for future steps ($j > t$).
- In rollout, predicted $\hat{z}$ is fed back into the model to predict further future states.

Key Formula 2: Training Objective and EMA Target

$$\mathcal{L}_{\mathrm{TF}} = \frac{1}{|\mathcal{V}|} \sum_{(b,t)\in\mathcal{V}} \left\| \hat{z}_{t+1}^{(b)} - z_{t+1}^{\mathrm{target},(b)} \right\|_1$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{TF}} + \mathcal{L}_{\mathrm{roll}}, \qquad \theta_{\mathrm{target}} \leftarrow \tau\theta_{\mathrm{target}} + (1-\tau)\theta_{\mathrm{online}}$$

- The target encoder provides a stop-gradient anchor to prevent the target space from moving erratically.
- Use ℓ_1 instead of MSE: it is more robust for heavy-tailed clinical variables.
- $\mathcal{L}_{\mathrm{roll}}$ trains multi-step stability with short autoregressive rollout.

Why Joint Training Can Fail

Failure 1: representation collapse

The encoder learns to output a constant, uninformative vector for every input ($z_{\text{std}} \approx 0$). The predictor can match it perfectly without learning any clinical dynamics.

Failure 2: online/target drift

As encoder parameters change, the latent space defined by the encoder also moves. The predictor learns to map from online space to target space; during autoregression, target-style outputs are fed back as online inputs, so errors amplify in a moving target space.

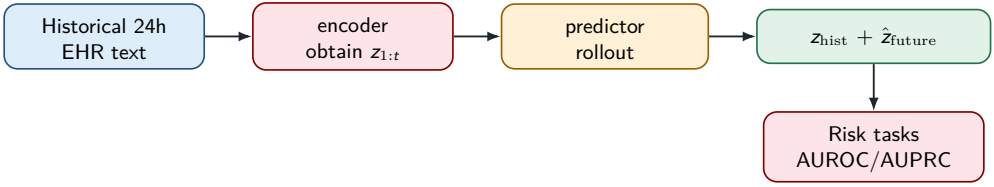
Clin-JEPA's contribution: a training curriculum that can stably retain the predictor.

Five-Stage Training Curriculum

This process uses an exponential moving average (EMA) target encoder (E_ξ) as a stable training anchor, with slowly updated target parameters ($\tau = 0.996$).

Stage	encoder LoRA	target encoder	Purpose
1 Warmup	frozen	online = target	First let the cold-start predictor learn basic trajectories and avoid collapse.
2 Co-training	trainable	EMA slow	The only stage where predictor gradients enter the encoder; the encoder learns embeddings specialized for predictor forecasting, i.e. dynamically grounded representations.
3 Alignment	frozen	EMA catches up	Let the target smoothly catch up with online and reduce space mismatch.
4 Hard sync	copy	target $\leftarrow$ online	Remove residual mismatch.
5 Finalize	frozen	online = target	The predictor further refines itself using its own outputs as feedback, improving inference-time stability.

How Clin-JEPA Is Used at Deployment



- z_{hist}: patient representation from observed history.
- $\hat{z}_{\text{future}}$: future latent state simulated by the predictor.
- Downstream tasks use the same backbone, without redesigning features for each task.

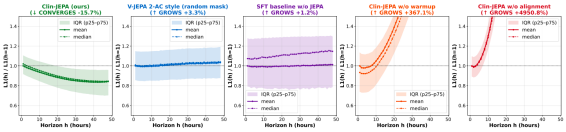
Experimental Setting: MIMIC-IV ICU

- Data: MIMIC-IV ICU, 84,497 stays and 64,874 patients.
- Split: patient-level 70/15/15 to avoid leakage from the same patient.
- Time: 1-hour bins, up to 72 hours; long stays use stride-12h windows.
- Training: 8 H200 GPUs, about 54 GPU-hours.

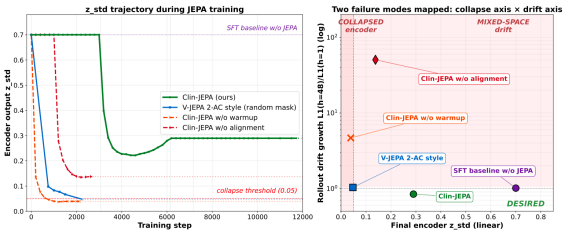
Why emphasize patient-level split?
If the same patient appears in both training and test sets, the model may memorize individual patterns and inflate generalization results.

Result 1: Only Clin-JEPA's Long-Horizon Rollout Converges

- Evaluate 48-hour rollout drift: the ℓ_1 distance between predicted latent states and true encoder outputs.
- Clin-JEPA: mean drift decreases by 15.7% relative to $h = 1$.
- Removing warmup: drift increases by 367%.
- Removing alignment: drift increases by 4951%.



(a) Error accumulation in predictor-simulated latent trajectories.

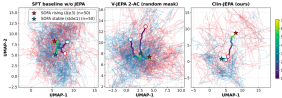


(b) Representation collapse and rollout drift as independent failure modes.

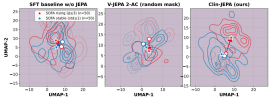
Result 2: Latent Space Separates Deteriorating and Stable Patients

- The paper selects two extreme phenotypes: deteriorating patients (increased severity, $\Delta\text{SOFA} \geq 3$) and stable patients.
- UMAP visualizes trajectories of hourly 4096-dimensional embeddings.
- In Clin-JEPA, latent displacement of deteriorating patients is 4.83 times that of stable patients.
- Baseline encoders show lower separation, with a maximum below 2.62 times.

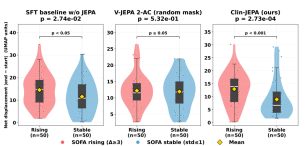
Interpretation: deterioration should correspond to larger latent movement.



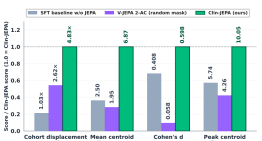
(a) **Trajectories.** Per-encoder trajectories with cohort means highlighted along a time-coded color band (hour 0→71).



(b) **Density contours.** Admission (dotted) vs. end-of-window (filled) cohort distributions; arrows mark mean displacement.

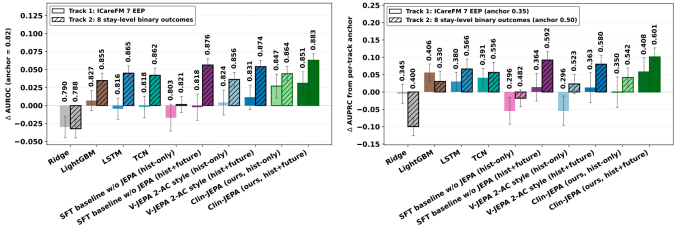


(c) **Per-patient net-displacement distributions.** One-sided Mann-Whitney U-test brackets and Cohen's d.



(d) **Four cross-encoder summary metrics**: cohort displacement ratio, mean centroid divergence, Cohen's d on net displacement, and peak centroid divergence.

Result 3: One Backbone Supports Multi-Task Risk Prediction



Evaluation track	Clin-JEPA	Meaning of improvement
7 ICareFM early event prediction tasks	AUROC 0.851	Stronger prediction of acute organ events.
8 hospital-level binary tasks	AUROC 0.883	Mortality and longer-term risk tasks benefit.
hist $\rightarrow$ full	Further gain	Future rollout adds information for risk tasks.

Figure: AUROC/AUPRC on two clinical benchmarks. This is retrospective evidence, not validation for real-time clinical decision-making.

Boundaries of Clin-JEPA

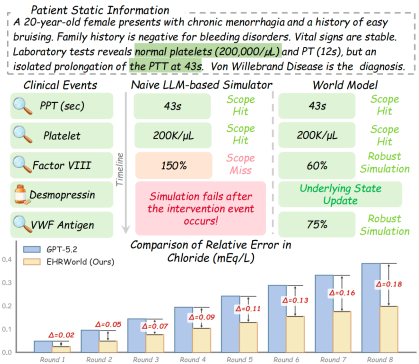
- It is a latent-space world model and does not directly generate readable laboratory values or images.
- The action is clinical intervention text recorded in EHR, not an identifiable intervention from a randomized trial.
- Training and evaluation focus on MIMIC-IV ICU; cross-hospital transfer still needs validation.
- The predictor improves risk tasks, but it is not a complete treatment-strategy optimizer.

Positioning:
Clin-JEPA is a latent predictive backbone for ICU EHR, showing that JEPA-style joint training can make LLM representations useful for patient-trajectory rollout.

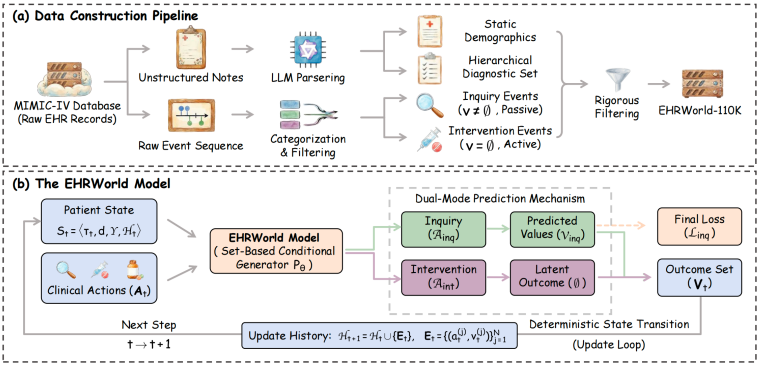
Outline

Question:
General medical LLMs know a lot of medical knowledge, but can they maintain a consistent patient state under continuous interventions?

- Static medical knowledge alone is not enough; long-horizon simulation can suffer from state inconsistency and error accumulation.
- EHRWorld-110K is constructed to train a patient-centered world model using temporal order and action outcomes.
- The output is not a single risk score, but future clinical events and observation results.



EHRWorld-110K Data Construction



Test set: 579 episodes, 84,010 inquiry events, and 25,798 intervention events.

Formalizing the Patient State

$$S_t = \langle \tau_t, d, Y, H_t \rangle$$

$$Y = \{\langle y_{\text{pri}}, r_{\text{pri}} \rangle\} \cup \{\langle y_{\text{sec}}^{(k)}, r_{\text{sec}}^{(k)} \rangle\}_{k=1}^{K}$$

- τ_t: current physiological timestamp.
- d: static information such as age and sex.
- Y: primary and secondary diagnoses with rationales, providing hospitalization context.
- H_t: complete interaction history from admission to the current time.

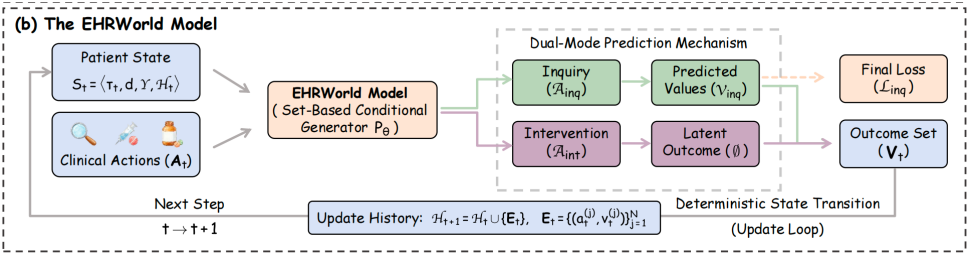
Set-Based Actions and Dual-Mode Prediction

$$A_t = \{a_t^{(1)}, \ldots, a_t^{(N)}\}$$

$$P_\theta(V_t | S_t, A_t) = \prod_{j=1}^{N} P_\theta(v_t^{(j)} | S_t, A_t)$$

- Multiple orders can be placed at the same clinical time point, so actions are represented as a set.
- Inquiry: tests/labs, for which the model predicts concrete observation values.
- Intervention: medications/procedures, which have no short-term output but are written into history and affect the future.

EHRWorld's State-Update Loop



$$E_t = \{(a_t^{(j)}, v_t^{(j)})\}_{j=1}^N,$$
$$S_{t+1} = \langle \tau_{t+1}, d, Y, H_{t+1} \rangle.$$

How EHRWorld Is Evaluated

Numeric variables

$$E_i = \left| \frac{y_i - \hat{y}_i}{y_i} \right|,$$

$$\text{S@}X = \frac{1}{N} \sum_i \mathbb{I}(E_i \leq X/100)$$

$$\text{SMAPE} = \frac{100\%}{N} \sum_i \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i| + \epsilon}$$

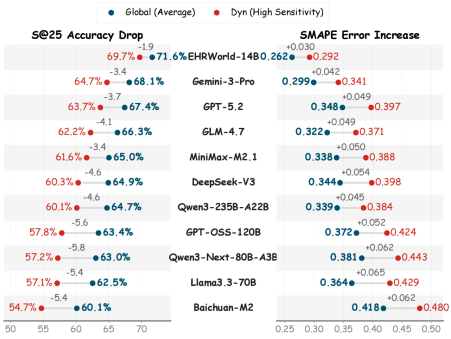
Clinical labels

- Clinical Status F1: whether normal/abnormal status is classified correctly.
- Label F1: discrete outcomes such as microbiology results.
- Retention Rate: fidelity of full trajectory relative to next-step prediction.

Main Results of EHRWorld

Model	S@25	SMAPE	Label F1
GPT-5.2	0.674	0.348	0.553
Gemini-3.0-Pro-Preview	0.681	0.299	0.477
EHRWorld-4B	0.703	0.274	0.912
EHRWorld-14B	0.716	0.262	0.913

- Strong LLMs still accumulate long-horizon errors.
- EHRWorld-14B achieves the best Avg Score of 0.765.
- Training built around time and action is more important than model scale alone; this is not strict causal identification.



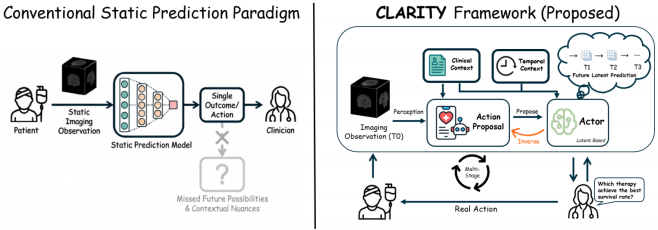
Boundaries of EHRWorld

- The focus is event trajectories over inpatient episodes, not an hourly ICU latent backbone.
- Interventions are written into history and affect future predictions, but this is not strict causal identification.
- Data come from MIMIC-IV; real deployment must face differences in coding, documentation habits, and treatment protocols across hospitals.
- Outputs are more readable, but long-horizon generation still needs clinical constraints, calibration, and uncertainty estimation.

Positioning:
EHRWorld shows that dedicated longitudinal EHR trajectory training can significantly improve long-horizon patient simulation.

Outline

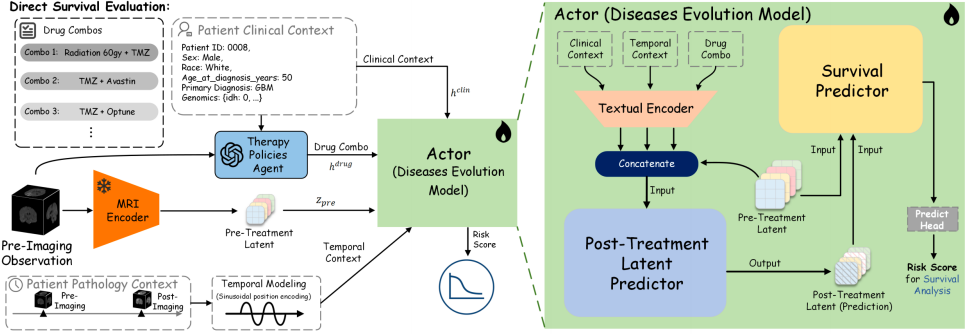
CLARITY v2: A World Model for Oncology Treatment Decisions



CLARITY's question:
Given pre-treatment MRI, patient clinical background, and candidate therapy, can we simulate the future disease trajectory and feed the simulation result back into treatment search?

- The v2 title emphasizes guiding treatment decisions by simulating context-aware disease trajectories.
- Data: MU-Glioma-Post, UCSF-ALPTDG external validation, and ISPY-2.
- The work moves from one-step prediction toward long-horizon prediction-to-decision.

CLARITY Inference Pipeline



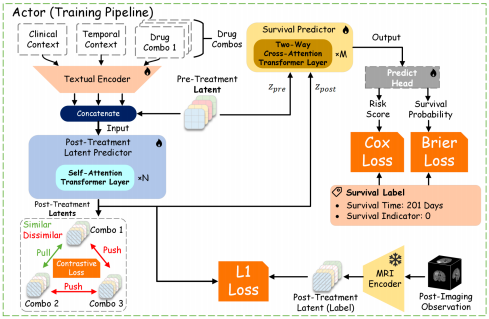
- The Therapy Agent proposes candidate regimens, and the Actor scores them one by one.
- Scores are fed back to the Agent, and the next round improves the regimen.
- This is closer to the clinical process of observing response and then adjusting therapy than one-shot static prediction.

CLARITY's Actor: Latent Disease Evolution

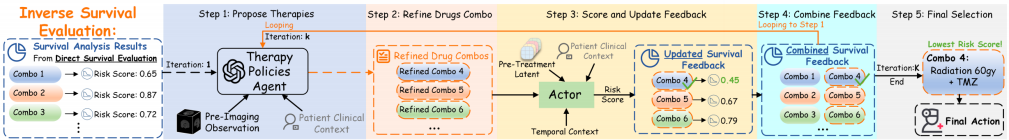
$$\hat{z}_{\text{post}} = \text{SelfAttn}\big([z_{\text{pre}}, h_{\text{clin}}, \phi(\Delta t), h_{\text{drug}}]\big)^N + z_{\text{pre}}$$

$$[\hat{p}_{1y}, \hat{r}] = \text{MLP}\big(\text{CrossAttn}(z_{\text{pre}}, \hat{z}_{\text{post}})^M + \text{CrossAttn}(\hat{z}_{\text{post}}, z_{\text{pre}})^M\big)$$

- h_{clin}: individualized context.
- $\phi(\Delta t)$: continuous-time encoding.
- h_{drug}: text encoding of the therapy regimen.
- Outputs include post-treatment latent, 1-year survival probability, and risk score.



Inverse Survival Evaluation: Turning Prediction into Decision



$$J(a_{1:H}) = \sum_{t=1}^{H} \beta^{t-1} \hat{r}_t$$

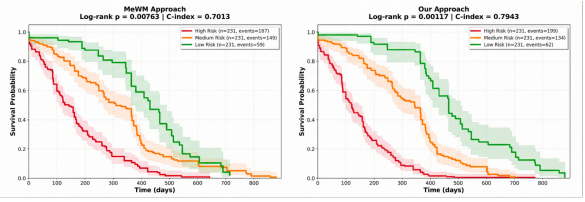
$$\min_{q} \ \mathbb{E}_{a_{1:H} \sim q}[J(a_{1:H})] - \tau \mathcal{H}(q)$$

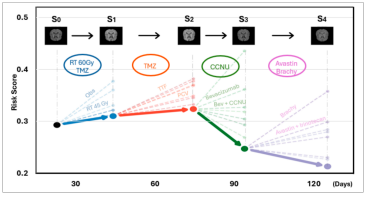
- Do not greedily choose a single action.
- Optimize cumulative risk over multi-step imagined trajectories.
- The entropy term encourages exploration and avoids locking onto a regimen too early.

CLARITY v2 Results and Boundaries

- MU-Glioma-Post: F1 57.1% vs. MeWM* 43.6%.
- UCSF-ALPTDG zero-shot: F1 50.2%.
- ISPY-2: F1 53.9%, extending to breast MRI.
- Survival analysis: C-index about 0.794.
- Latent inference: about 0.341s, faster than diffusion sampling.

Boundary: retrospective validation; not an automatic order-entry system.





Outline

Similarities and Differences Across the Three Models

Dimension	Clin-JEPA	EHRWorld	CLARITY v2
Main data	Hourly ICU EHR text	Full-stay inpatient EHR events	MRI + clinical/genomic + treatment
State repr.	4096-d patient latent	Explicit state $S_t = \langle \tau, d, Y, H \rangle$	imaging-derived latent
Action	Hourly intervention text	inquiry/intervention action set	therapy schedule / drug combo
Dynamics	Retained predictor for latent rollout	Conditional generator for event outcomes	Actor predicts post-treatment latent
Output	Future latent + risk features	Lab values, state labels, discrete outcomes	latent trajectory + survival risk
Main contribution	Stable co-training curriculum	Longitudinal EHR data and long-horizon simulation	Prediction-to-treatment feedback loop
Main boundary	Latent space is not directly readable; causality is weak	Readable generation, but still observational data	Depends on therapy agent and rule constraints

Take-Home Messages and References

1. The core of a medical world model is dynamic transition: state + action $\rightarrow$ future.
2. Clin-JEPA turns JEPA into an EHR latent rollout backbone.
3. EHRWorld emphasizes long-horizon EHR event simulation, showing that dedicated dynamic training is more important than static medical knowledge alone.
4. CLARITY emphasizes treatment-conditioned latent trajectories and survival-feedback planning.
5. None of the three is an automatic clinical decision system; causal validation, safety constraints, and cross-institution generalization remain central.

Main references

- Yang et al. Clin-JEPA, arXiv:2605.10840v2, 2026.
- Mu et al. EHRWorld, arXiv:2602.03569v1, 2026.
- Ding et al. CLARITY, arXiv:2512.08029v2, 2026.
- Ha and Schmidhuber. World Models, 2018.
- Hafner et al. PlaNet, 2019.
- Assran et al. I-JEPA, 2023; V-JEPA 2, 2025.
- Johnson et al. MIMIC-IV, PhysioNet / Scientific Data.
- Qazi et al. Beyond Generative AI: World Models for Clinical Prediction, Counterfactuals, and Planning, 2025.