CLINTIC: LEARNING CLINICAL DYNAMICS THROUGH TREATMENT-CONDITIONED NEURAL SDES

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose CLINTIC, a foundational model paradigm for *clin*ical trajectories built on a treatment-conditioned latent neural stochas*tic* differential equation (SDE). Patient state evolves continuously between sparse, irregular events with treatment as the driving action; CLINTIC encodes multi-modal EHR streams into the SDE latent under Markov decision process (MDP)-aligned action conditioning, and per-step decoders read out clinically meaningful targets across the rollout. An IsoFLOP sweep over seven depths and six budgets (3×10^{12} to 10^{15} FLOPs) traces a clean Chinchilla-form frontier with size exponent $N^*(C) \propto C^{0.25}$, showing the paradigm is also scalable, with data more limiting than parameter count at this scale. CLINTIC outperforms existing EHR foundation models and frontier LLMs across heterogeneous real-world cohorts, with the largest gain on a RECIST-PD-analogue cancer progression target (AUROC 0.827, $+0.115$ over Opus-5). Risk representations read directly off CLINTIC's dynamics separates 52-week OS where no standard biomarker reaches significance (HR 3.27 overall, 6.35 on Nivolumab); a further regimen-substitution probe on the same dynamics yields a stratifier of comparable discriminative power. Codes are available at `https://anonymous.4open.science/r/ClinTic-3807`.

1 INTRODUCTION

Real-world clinical trajectories are long, irregularly sampled, multi-modal, and high-stakes. Continuous labs and vitals, MD-authored performance and disease assessments, free-text pathology and imaging narratives, genomic profiles, and irregular treatment events coexist in a single patient record (Rajkomar et al. (2018); Harutyunyan et al. (2017); Johnson et al. (2023)). Most predictive models cast this as static classification: given features at a fixed time, predict an outcome. This discards temporal structure and cannot answer counterfactual questions clinicians and trialists pose: what would have happened on a different regimen, would this patient meet a synthetic-control criterion at month six, what is the expected ICU trajectory under a different ventilation strategy.

World models (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2023) learn latent simulators that roll forward under any chosen action, and recent clinical adaptations apply the idea to sepsis decision support (Xu et al., 2026) and counterfactual treatment-effect estimation (Li et al., 2021; Melnychuk et al., 2022; Bica et al., 2020). Each is bespoke per cohort and modality mixture, leaving no general framework whose continuous-time dynamics match the biology and whose ingestion absorbs heterogeneous EHRs without recipe-level rework. Recent EHR-transformer scaling studies (Zhang et al., 2025; Waxler et al., 2025) report Chinchilla-form fits on tokenized next-event prediction; the corresponding question for treatment-conditioned generative trajectory dynamics, the regime CLINTIC targets, has not been measured. We aim to close that gap with CLINTIC, instantiated on **MSK CCDE** (oncology, $23{,}320$ patients, weekly bins) and **MIMIC-IV** (Johnson et al., 2023) (ICU, $68{,}546$ stays, hourly bins) under one architecture and training recipe.

Contributions. (i) *A scalable framework for clinical trajectory modeling* built on a treatment-conditioned latent SDE with event-level multi-modal fusion and MDP-aligned action conditioning, absorbing heterogeneous EHRs without per-cohort rework. (ii) *Scaling-law investigation.* An IsoFLOP sweep (3×10^{12} to 10^{15} FLOPs, seven depths) traces a Chinchilla-form frontier with size exponent $N^*(C) \propto C^{0.25}$ (95% CI $[0.16, 0.34]$), showing the paradigm scales predictably with data

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.

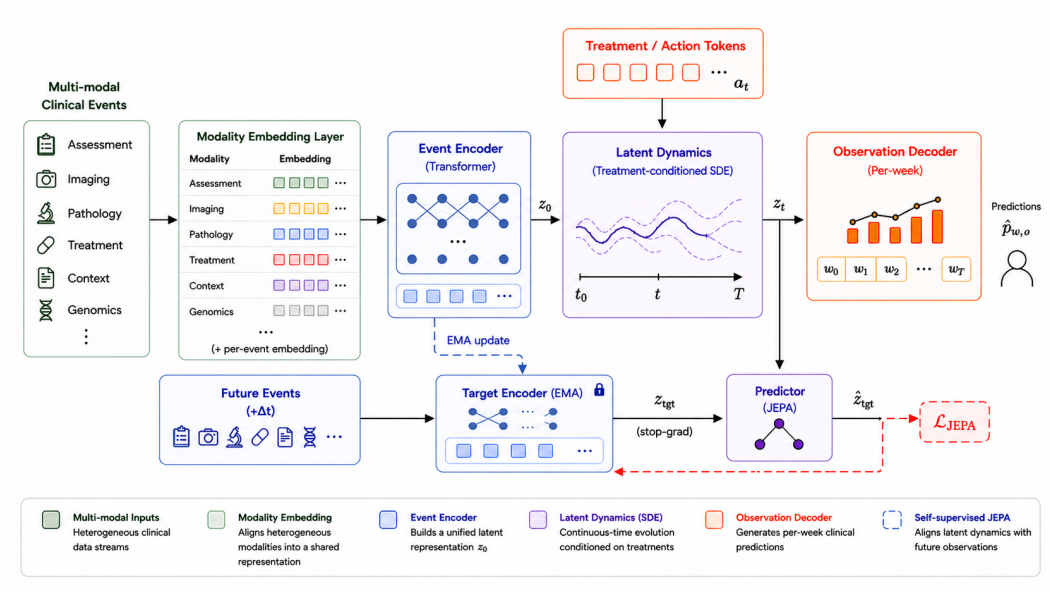


Figure 1: Architecture overview (t_d=dream start). **Top:** the encoder ingests different modality streams from the encode window $[t_{\text{first}}, t_d)$ and emits K=16 readout tokens $\mathbf{z}_0 \in \mathbb{R}^{K \times d_z}$; a treatment-conditioned neural SDE rolls forward W bins under interventions $\mathbf{a}_t$; the per-step observation decoder produces cohort-specific outputs from each $\mathbf{z}_t$. **Bottom (JEPA):** an EMA encoder copy (gradients disabled) encodes the dream-window events into $\mathbf{z}_{\text{tgt}}$; a lightweight MLP predicts $\hat{\mathbf{z}}_{\text{tgt}}$ from the time-averaged dreamed latents; SmoothL1 on ℓ_2-normalized vectors. Appendix D expands the per-modality fusion and SDE rollout.

breadth, not parameter count, as the binding constraint at this scale. (iii) *Cross-cohort evidence.* CLINTIC wins all five MSK CCDE targets against 13 baselines and all three MIMIC-IV stay-level outcomes against CLMBR-T, ETHOS, EHRMamba (+0.066 AUROC, +0.189 AUPRC, +0.063 LOS accuracy). (iv) *Patient-level risk representations.* The dynamics yield representations encoding patient-level prognostic information for downstream clinical tasks: a landmark-KM score off the dream window separates 52-week OS on MSK CCDE (HR 3.27 overall, 6.35 on Nivolumab) without a survival head; an observational counterfactual regimen-substitution probe on the same dynamics gives a comparable stratifier on NSCLC 52-week OS (HR 4.59 $[3.14, 6.70]$ on n=1,108 NSCLC under Carboplatin/Paclitaxel; factual landmark baseline HR 3.04 $[2.08, 4.44]$ on the same cohort).

2 METHOD

CLINTIC has four components (see Figure 1): per-modality event encoders, a bidirectional Transformer (Vaswani et al., 2017) event encoder mapping patient history into K readout tokens, a treatment-conditioned latent SDE dynamics module that rolls forward W steps at a cohort-configured cadence, and an observation decoder producing per-step outputs (categorical heads, continuous channels, or both). A JEPA target encoder provides a self-supervised auxiliary loss. Cohort instantiation amounts to selecting the modality registry, step cadence, and decoder head structure (§3.1); Appendix D expands per-modality fusion and action–latent interactions.

2.1 EVENT ENCODER

Per-modality embeddings. Each event maps to a d_e-dim vector $\mathbf{e}_i = \text{Enc}_{\tau_i}(\text{raw}_i) + W_{\text{mod}}[\tau_i]$; the $W_{\text{mod}}[\tau_i]$ tag preserves channel routing in the shared embedding space. Sub-encoders are pluggable: text-serialisable events flow through a clinical LM encoder (mean-pool to d_e); tabular modalities through small MLPs; any encoder emitting a d_e-dim vector plugs in without downstream changes. Per-cohort choices: §3.1.

Continuous-time RoPE. Because events are irregularly sampled, we replace integer position indices with raw timestamps t_i in cohort-natural units inside rotary position embedding (Su et al., 2024): $\theta_k(t) = t \cdot 10000^{-2k/d_h}$. The sequence is prepended with K learnable readout tokens at a fixed offset before the dream window and an always-visible demographics anchor.

Transformer. A bidirectional Transformer based event encoder attends jointly over readout, context, and event positions. We adopt Flash Attention 3 (Shah et al., 2024) to compute attention over batches packed without padding. The output is $\mathbf{z}_0 = W_{\text{out}}\mathbf{r} \in \mathbb{R}^{K \times d_z}$; depth and width are cohort-tuned within the IsoFLOP envelope of §3.4.

2.2 Treatment-conditioned neural SDE

Given $\mathbf{z}_0$, dynamics evolve over W bins with S Euler–Maruyama sub-steps per bin (Oksendal, 2013; Kidger et al., 2021; Li et al., 2020a):

$$d\mathbf{z}_t = f_\theta(\mathbf{z}_t, \mathbf{a}_t, t)\, dt + \sigma_\theta(\mathbf{z}_t)\, d\mathbf{W}_t, \tag{1}$$

where $\mathbf{a}_t$ is the action token (mean-pool of treatment-event embeddings within the bin, or a learned no-treatment placeholder), f_θ a Transformer drift, σ_θ a state-dependent softplus diffusion. The drift attends over $[\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(K)}, \mathbf{a}_t]$ but updates only the K latent tokens; the action acts as conditioning, not state.

Treatment enters the dynamics as an external conditioning input rather than a state variable to be generated. This separates the supplied treatment sequence from the modeled patient trajectory and permits rollouts under different action sequences. However, treatment assignment in observational EHRs is confounded; this architectural factorization alone does not identify causal treatment effects. State-dependent diffusion introduces stochastic variation in latent trajectories, whose contribution we evaluate against a treatment-conditioned deterministic ODE ablation.

Temporal ordering of actions and predictions. The per-step decoder reads $\mathbf{z}_t$ *before* treatment $\mathbf{a}_t$ is applied,

$$\text{traj}[t] = \mathbf{z}_t, \qquad \mathbf{z}_{t+1} = \mathbf{z}_t + f_\theta(\mathbf{z}_t, \mathbf{a}_t, t)\, \Delta t + \sigma_\theta(\mathbf{z}_t)\sqrt{\Delta t}\, \boldsymbol{\varepsilon}. \tag{2}$$

This prevents same-step treatment events, which often reflect outcomes (e.g. stopping a drug because of progression), from leaking into the corresponding observation. The dynamics are therefore autoregressive in treatment: prediction at step T depends only on actions $0..T{-}1$, the property exploited in §3.6 for both landmark risk scoring and the regimen-substitution counterfactual probe.

2.3 Decoder, JEPA, and training

Decoder. A small MLP maps $\mathbf{z}_t$ to per-step outputs: a multi-head categorical decoder for discrete chart fields (cross-entropy) and/or a multi-channel continuous decoder for physiological streams (Smooth L1). Targets bin ground-truth events into the W dream bins with last-write-wins aggregation; bins without observations are dropped from the loss. Stay-level outcome heads (mortality, LOS) read from the final dream latent when the cohort defines them.

State regularization via JEPA. A fundamental challenge in EHR modeling is that the true patient state is often *under-determined* by structured measurements; the most critical clinical nuances which found in pathology narratives or diagnostic reasoning cannot always be mapped to a discrete decoder head. To capture this unquantified information, we incorporate a Joint-Embedding Predictive Architecture (JEPA) (Assran et al., 2023) branch that regularizes the latent trajectory without requiring explicit decoding. An EMA copy of the encoder encodes dream-window events into $\mathbf{z}_{\text{tgt}}$; a lightweight MLP predicts $\hat{\mathbf{z}}_{\text{tgt}}$ from the time-averaged dreamed latents and we apply SmoothL1 on ℓ_2-normalized vectors. JEPA adds a dense self-supervised signal not bound by decoded-target sparsity, intended as a representation-shaping auxiliary on patient-state targets. §3.5 bears this out. We minimize $\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{JEPA}}$.

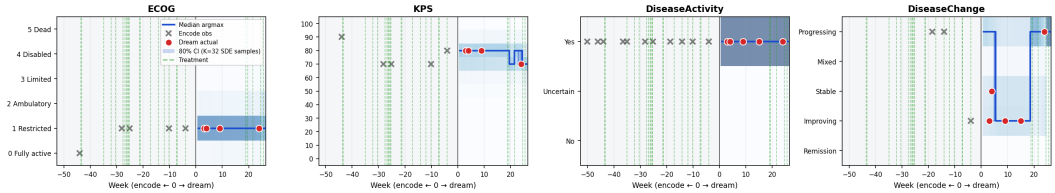


Figure 2: Per-observation forecasts for a heavily-treated test patient (60 encode events, 20 treatments). Median argmax (blue step), mean class-probability heatmap, and 80% CI band from $K{=}32$ stochastic SDE rollouts (10–90 pct of per-sample argmax). Red dots: dream-window ground truth; grey $\times$: encode-window observations; dashed green: treatment events. The CI band narrows where confident (DiseaseActivity flat at *Yes*, $6/6$; ECOG at *Restricted*, $5/5$) and widens at multi-modal transitions (DiseaseChange around weeks 5 and 18, $3/5$ correct; KPS late dip $80{\to}70$, $4/5$). Latent dynamics for the same patient and horizon stability across the test split are in Appendix F.

3 Experiments

We organise experiments by cohort. §3.2–§3.6 report on **MSK CCDE** (oncology, weekly): setup, twelve external baselines, IsoFLOP scaling, and a leakage-free landmark KM analysis. §3.7 reports cross-cohort validation on **MIMIC-IV** (Johnson et al., 2023) (ICU, hourly). Both cohorts share the backbone, JEPA branch, optimizer, and schedule from §2; only modality registry, observation targets, and step cadence vary per cohort.

3.1 Cohort instantiations

MSK CCDE. 23,320 patients (80/20 patient-level split): $\sim$74 events per patient (median) across 31 source tables and 6 modalities, along a timeline whose median reach is age $\sim$70 y. Six modalities (assessment, imaging, pathology, treatment, genomics, context) feed five MD-authored chart fields decoded at $W{=}26$ weekly steps over a 182-day dream window: ECOG (Oken et al., 1982), KPS (Karnofsky, 1949), disease state, disease activity, and disease change (5-class RECIST-analogue (Eisenhauer et al., 2009)). Genomics events use a dedicated MLP over a $541{\times}6$ gene-by-channel matrix; other modalities are serialized as header-and-rows text and tokenized with ClinicalBERT (Alsentzer et al., 2019). Full distribution statistics: Appendix B (Table 5).

MIMIC-IV v3.1 ICU. 68,546 ICU stays under inclusion criteria {adults ($\geq$ 18 y), first ICU admission, total stay $\geq$ 24 h}; 80/20 stay-level split. Five modalities (vitals, labs, neuro/output, treatments, context) feed two dream-stage heads: 68 continuous physiological channels decoded at $W{=}24$ hourly steps (Smooth L1), and stay-level mortality + 10-class length-of-stay (Harutyunyan et al., 2017) read from the final dream latent. Side-by-side instantiation and modality/target registry: Appendix B; data vocabulary and event serialization: Appendix C.

3.2 Setup

The model emits a dense $26{\times}5$ grid of class probabilities per test patient. We score only cells with a ground-truth weekly observation, using the same last-write-wins binning as training, and report patient-level bootstrap CIs (200 resamples over patients) to preserve within-patient trajectory correlation. Primary metrics track class structure and clinical decision: ECOG and KPS (ordinal performance) use quadratic-weighted κ (QWK); DiseaseState (extent: localized$\to$distant) uses AUROC@($y{\geq}2$), the metastatic threshold; DiseaseActivity (nominal) uses macro-F1; DiseaseChange (RECIST-analogue) uses AUROC@($y{\geq}4$), the progressive disease threshold. Threshold AUROC ranks classes consistently with the clinical decision they drive without committing to a calibration. Appendix E gives full metric definitions and per-obs rationale.

Figure 2 grounds the per-cell evaluation in a single patient: the trajectory view overlays weekly argmax predictions on ground-truth observations across the five chart-field heads. Latent L2-drift impulses concentrate at treatment weeks rather than co-occurring uniformly (Figure 6, appendix), and primary metrics are approximately flat across the 26-week horizon with no long-horizon collapse

Table 1: Comparison against twelve external baselines on the 4,664-patient test split, identical (patient, week, obs) cells. Each cell: mean with 95% patient-level bootstrap CI; ECOG/KPS use QWK, disease state (DSt) AUROC@2, disease activity (DAct) macro-F1, disease change (DCh) AUROC@4; higher is better; **bold** = column max.

Method	ECOG	KPS	DSt	DAct	DCh
Uniform	0.000 [.000,.000]	0.000 [.000,.000]	0.500 [.500,.500]	0.146 [.139,.152]	0.500 [.500,.500]
Most-frequent	0.000 [.000,.000]	0.000 [.000,.000]	0.500 [.500,.500]	0.271 [.267,.274]	0.500 [.500,.500]
LogReg	0.342 [.312,.370]	0.272 [.235,.302]	0.956 [.943,.967]	0.496 [.483,.506]	0.666 [.644,.689]
XGBoost	0.301 [.273,.330]	0.242 [.209,.272]	0.950 [.937,.962]	0.527 [.512,.544]	0.646 [.624,.669]
CRN	0.368 [.331,.404]	0.303 [.259,.338]	0.951 [.938,.964]	0.528 [.510,.545]	0.671 [.651,.692]
Latent ODE	0.348 [.307,.385]	0.324 [.289,.359]	0.953 [.939,.965]	0.555 [.536,.578]	0.693 [.673,.712]
Dynamic-DeepHit	0.351 [.315,.383]	0.288 [.245,.323]	0.949 [.935,.962]	0.523 [.504,.543]	0.684 [.662,.703]
G-Net	0.061 [.046,.080]	0.044 [.025,.065]	0.920 [.901,.937]	0.384 [.375,.393]	0.645 [.626,.668]
Causal Transformer	0.360 [.328,.391]	0.289 [.253,.322]	0.953 [.937,.966]	0.514 [.500,.528]	0.702 [.681,.724]
Claude Sonnet-4.6	0.209 [.185,.237]	0.130 [.107,.154]	0.971 [.962,.980]	0.408 [.398,.416]	0.685 [.664,.704]
Claude Opus-4.7	0.333 [.305,.366]	0.192 [.160,.233]	0.973 [.962,.984]	0.484 [.464,.502]	0.703 [.680,.728]
Claude Opus-5	0.342 [.314,.375]	0.203 [.171,.244]	0.979 [.970,.992]	0.495 [.475,.513]	0.712 [.689,.737]
CLINTIC	**0.437** [.408,.472]	**0.396** [.361,.436]	**0.979** [.971,.987]	**0.605** [.587,.626]	**0.827** [.810,.844]

on the high-data observations (Figure 7, appendix). Unless stated otherwise, all comparisons below use the depth-6 production cell, justified compute-optimally in §3.4.

3.3 COMPARISON AGAINST EXTERNAL BASELINES

We benchmark against twelve baselines on identical cells (patient, week, obs): ten classical and representation-learning baselines (rule-based floors, tabular ML, RNN/Transformer counterfactual sequence models, a neural ODE, a longitudinal survival model) and frontier LLMs Claude Sonnet-4.6 and Opus-4.7 and Opus-5. Trained baselines see the same train split as CLINTIC; the LLMs see each patient's pre-prediction window as a chronological event log (relative day offsets only) with demographics and a gene-list summary, emitting per-week class distributions in a strict JSON schema. Appendix G lists configurations and the LLM prompt/parser. Table 1 reports per-obs primary metrics with patient-level 95% bootstrap CIs.

3.4 SCALING LAWS

We sweep the depth of model $L \in \{2, 3, 4, 6, 8, 12, 16\}$ and the FLOP budget $F \in \{3{\times}10^{12}, 10^{13}, 3{\times}10^{13}, 10^{14}, 3{\times}10^{14}, 10^{15}\}$ as an 7×6 IsoFLOP grid, with all hyperparameters derived from depth via a single dial: $d_e = \mathrm{round}_{16}(128L/6)$, $d_z = 2d_e$, $d_{\text{ff}} = 4d_e$, $L_{\text{dyn}} = \max(1, \lfloor L/3 \rfloor)$, $n_{\text{heads}} = d_z/32$. Each cell trains to its target FLOP budget; we evaluate the best validation checkpoint with the dense per-cell suite (Table 9, appendix).

Compute-optimal frontier. The compute-optimal envelope (per-budget min over $L \in \{2, 3, 4, 6, 8\}$; the two deepest cells $L \in \{12, 16\}$ never reach the envelope at any budget in the sweep) walks $L{=}2{\to}3{\to}4{\to}6$ as compute grows (Figure 3). It fits a Chinchilla-form $L^*(C) = E + bC^{\beta}$ with $\hat{E}{=}0.646$, $\hat{b}{=}1.73{\times}10^6$, $\hat{\beta}{=}{-}0.568$ at RMS residual 0.004, and $\hat{N}^*(C) = 1.05{\times}10^3\, C^{0.250}$.

Residual-bootstrap on the envelope (5,000 resamples) gives 95% CIs $\hat{E} \in [0.635, 0.653]$, $\hat{\beta} \in [-0.685, -0.464]$, $\hat{\alpha} \in [0.164, 0.336]$; leave-one-budget-out re-fits give $\alpha \in [0.16, 0.30]$ and $\beta \in [-0.70, -0.51]$, with the size exponent below the Chinchilla $\alpha{=}0.50$ and Kaplan $\alpha{=}0.73$ in every resampling. The continuous-time SDE training objective on irregular EHR streams traces a Chinchilla-form frontier at a fit quality that lets us read compute-optimal model size and data exponents directly off the envelope; the recovered $\alpha = 0.25$ also sits well below the $\alpha \approx 0.58$ (N^*-exponent) reported by Zhang et al. (2025) and the $\alpha = 0.520$ reported by Waxler et al. (2025) for tokenized next-event EHR transformers. The IsoFLOP methodology is matched, but cohort, training objective, and tokenization all differ across these reports, so the exponent gap locates this measure-

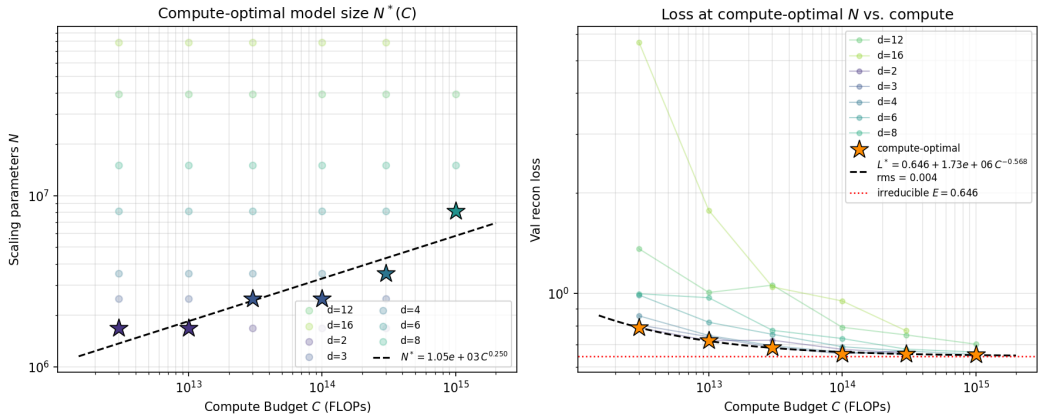


Figure 3: Compute-optimal scaling. *Left*: $N^*(C) = 1.05{\times}10^3\, C^{0.250}$ (dashed); envelope cells colour-coded by depth. *Right*: Chinchilla form $L^*(C) = 0.646 + 1.73{\times}10^6\, C^{-0.568}$ (dashed, RMS residual 0.004); irreducible-loss asymptote $E = 0.646$ in red. Bootstrap 95% CIs: $E \in [0.64, 0.65]$, $\beta \in [-0.69, -0.46]$, size exponent $\alpha \in [0.16, 0.34]$.

Table 2: Component ablations of CLINTIC at L=6. Point estimates only; CIs match Table 1. **Bold** = column max.

Variant	ECOG	KPS	DSta	DAct	DChg
CLINTIC	**.437**	**.396**	**.979**	**.605**	**.827**
− JEPA	.416	.355	.972	.579	**.827**
− SDE (σ=0)	.420	.387	.970	.602	.812
− action	.419	.361	.972	.579	.738

ment (continuous-time SDE on irregular multi-modal EHR, MSK CCDE) in a more data-bound regime than tokenized next-event reports rather than identifying the SDE backbone alone as the cause. Though the small exponent reflects efficient compute use rather than weak scaling: downstream gains continue with budget along the envelope (Table 9), and at 10^{15} FLOPs the production cell already exceeds twelve external baselines (§3.3). We use depth-6 (∼78M total parameters; 8.2M non-embedding N for the scaling fit, with frozen ClinicalBERT word-piece embeddings excluded following Kaplan et al. (2020); Hoffmann et al. (2022)) for all other evaluations. Operating-point analysis, downstream-metric tracking on and off the envelope, and caveats on the fit are deferred to Appendix H.

3.5 COMPONENT ABLATIONS

We isolate the contribution of three design choices by retraining CLINTIC at the production depth (identical settings) under three single-knob ablations. **(i) no JEPA** disables the auxiliary latent-prediction loss and the EMA target encoder, leaving only per-week reconstruction. **(ii) no SDE** zeroes the diffusion coefficient ($\sigma \equiv 0$), collapsing the latent rollout to a deterministic ODE while keeping the same compute. **(iii) no action** replaces every per-week action token with a constant null embedding, so treatment events still appear in the encoder as passive history but no longer condition the dynamics.

The three knobs separate cleanly. Removing action conditioning collapses DiseaseChange (CIs disjoint with the full model) and degrades DiseaseActivity, the two targets that should reflect treatment effect, while DiseaseState (metastatic vs. not, largely fixed within the dream window) stays within the full-model CI; this directly validates treatment conditioning as the carrier of the change-vs.-stable signal and motivates the leakage analysis in §3.6. Removing JEPA hurts patient-state ordinals selectively but matches the full model on DiseaseChange, consistent with the auxiliary representing *which state* a patient is in rather than *how state evolves under treatment*. The −SDE ablation is the controlled treatment-conditioned ODE within our framework, isolating stochasticity from every other design choice that distinguishes CLINTIC from Latent ODE; on point-estimate accuracy it is

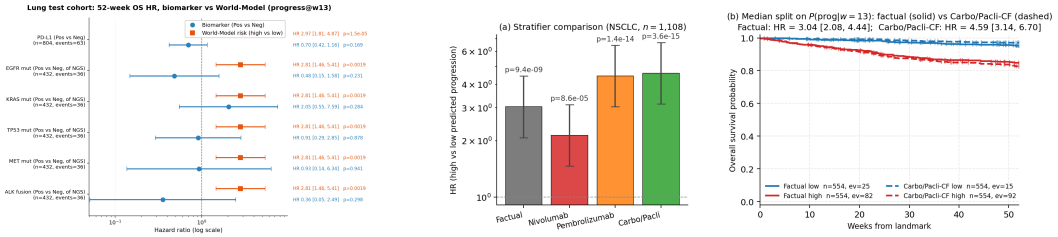


Figure 4: **Left:** Univariate landmark-KM HRs for s_{13} versus six standard NSCLC biomarkers (PD-L1, EGFR, KRAS, TP53, MET, ALK) on the lung test subset, all run through an identical pipeline; only s_{13} reaches significance. **Right:** Regimen-counterfactual probe on the NSCLC test cohort (n=1,108, 107 events at 52 w): every dream-window action token is replaced by the learned regimen-R token (mean of ACTIONPROJ(EVENTEMB(e)) over factual R events), giving $s_i^R = P_i^R(\text{prog})|_{w=13}$, the patient's progression forecast under a counterfactual where they receive R throughout the dream. **(a)** Median-split landmark-KM HR(high vs. low) for the factual baseline s_{13}^{fact} (gray) and three counterfactual regimens on the same cohort. **(b)** KM curves under median split on $P(\text{prog})|_{w=13}$: factual (solid) vs Carbo/Pacli-counterfactual (dashed) on the same cohort.

Table 3: Per-cohort landmark KM HRs at week T on the median split of s_T. All test log-rank $p < 10^{-300}$; Nivolumab T=13 p=7.1×10^{-7}.

Cohort	T	n / ev.	HR (95% CI)
All test	1	4393/374	3.74 [3.05, 4.58]
All test	4	4369/371	3.83 [3.12, 4.69]
All test	13	4285/374	3.27 [2.67, 4.01]
Nivolumab	1	89/31	1.94 [0.96, 3.93]
Nivolumab	4	89/32	3.59 [1.78, 7.26]
Nivolumab	13	85/30	**6.35 [3.06, 13.17]**
Pembrolizumab	4	70/21	2.40 [1.01, 5.67]
Carbo/Pacli.	4	156/12	5.60 [1.80, 17.38]
FOLFIRINOX	4	86/25	2.27 [1.03, 5.02]

the mildest knob, but multi-class Brier favors the full SDE on every observation (Appendix J), so the learned diffusion $\sigma_\theta(\mathbf{z}_t)$ supplies probabilistic refinement the controlled ODE cannot.

3.6 LANDMARK RISK STRATIFICATION WITH COUNTERFACTUAL SIMULATIONS

We ask whether the trained dynamics already supply a patient-level risk score as a derived statistic, without a survival head or stratification-specific target. The biomarker comparison above is evidence that the framework learns representations carrying prognostic information beyond what any single standard NSCLC biomarker contributes in this cohort, not a head-to-head against a properly tuned multivariate survival model. A natural choice, $\max_t P(\text{DChg=progressing})$ over the 26-week dream, gives strong-looking KM separation that collapses (log-rank p: $8.8\times10^{-6}\to0.15$) under a no-treatment ablation replacing every action with the learned placeholder, indicating that future-treatment tokens leak into the prediction (drug-end events frequently encode progression). We treat this no-treatment ablation as a required diagnostic for any treatment-conditioned trajectory model claiming leakage-free prognostic value, and exploit Eq. 1: the prediction at week T depends only on actions for weeks $0..T{-}1$. We define $s_T = P(\text{DChg=progressing})$ at a single landmark T and shift the KM origin to dream_start + $7T$ d so the curves describe trajectory *beyond* the landmark. Patients with death_time $\leq$ landmark are excluded; deaths within 52 w of the landmark are events, others censored at $\min(\text{last_event} - \text{landmark}, 52\,\text{w})$. We report $T \in \{1, 4, 13\}$ (T=13 matches the 3-month clinical assessment landmark).

Results. Stratifying on the median of s_{13} across the full test split gives HR $3.27\,[2.67, 4.01]$ ($p \ll 1e^{-3}$, n=4,285, 374 events); on named-regimen cohorts (Table 3) Nivolumab reaches HR $6.35\,[3.06, 13.17]$ at T=13 with a monotone trend in T (HR 1.94→3.59→6.35 at T=1→4→13), consistent with more treatment-conditioned dynamics running autoregressively before the readout; T trades signal for power as cohorts attrit. On the lung test cohort, s_{13} is the only stratifier reach-

Table 4: MIMIC-IV stay-level outcomes on the 13,721-stay held-out test split. Each cell: mean with 95% stay-level bootstrap CI (1000 resamples); higher is better; **bold** = column max.

Method	Mort. AUROC↑	Mort. AUPRC↑	LOS Acc↑
Prior	0.500 [.500,.500]	0.112 [.107,.117]	0.394 [.386,.402]
CLMBR-T (Steinberg et al., 2020)	0.708 [.694,.722]	0.231 [.215,.248]	0.421 [.412,.429]
ETHOS (Renc et al., 2024)	0.832 [.823,.844]	0.452 [.422,.463]	0.479 [.471,.488]
EHRMamba (Fallahpour et al., 2024)	0.851 [.840,.862]	0.476 [.456,.497]	0.491 [.482,.500]
CLINTIC	**0.917** [.911,.923]	**0.665** [.652,.679]	**0.554** [.545,.562]

ing significance against six NSCLC biomarkers (Figure 4 top: PD-L1, EGFR, KRAS, TP53, MET, ALK; HR $\sim$2.8–3.0, $p \ll 1e^{-3}$), and the PD-L1$\times s_{13}$ cross-tabulation is only weakly associated (53% vs. 48%), so the scores carry largely orthogonal information.

Counterfactual regimen substitution. With $\mathbf{a}_t$ exogenous, the trained SDE supports counterfactual rollouts by construction: any patient's factual action stream can be swapped for an alternative and the dynamics re-integrated. We use the simplest informative intervention, holding $\mathbf{a}_t$ at a chosen regimen for the entire dream window. For $R \in \{\text{Nivolumab, Pembrolizumab, Carbo/Pacli}\}$ on the NSCLC test cohort (n=1,108, 107 events at 52 w), replacing every dream-window action token with the learned regimen-R token (the mean of $\textsc{ActionProj}(\textsc{EventEmb}(e))$ over factual R events) yields $s_i^R = P_i^R(\text{prog})|_{w=13}$, the model's progression forecast for patient i under a counterfactual where they receive R throughout the dream. Median split on s_i^R separates 52-week landmark OS with HR(high vs. low) $2.14\,[1.46, 3.12]$ (Nivolumab), $4.44\,[3.04, 6.48]$ (Pembrolizumab), and $4.59\,[3.14, 6.70]$ (Carbo/Pacli; Figure 4b). The factual stratifier s_{13}^{fact} on the same NSCLC cohort gives HR $3.04\,[2.08, 4.44]$ (Figure 4a, gray bar), so the counterfactual stratifiers reach discriminative power comparable to the factual baseline despite reading from a substituted action stream rather than the patient's actual treatment history. Crucially, because observational EHR data lacks a real counterfactual ground truth, these results should be interpreted as diagnostic proxies for model coherence rather than definitive causal estimates of treatment effect with full counterfactual identification under confounding by indication left to follow-up.

3.7 CROSS-COHORT VALIDATION: MIMIC-IV

We retrain CLINTIC on MIMIC-IV (§3.1) using the identical backbone, JEPA branch, optimizer, and schedule from §2.3; only the modality registry, per-step decoder head (68 Smooth-L1 channels in place of five categorical heads), stay-level outcome heads (binary mortality, 10-class LOS), and step cadence (W=24 hourly bins) differ. Per-stream reconstruction is scored by Smooth L1 on observed (stay, hour, channel) cells of the 13,721-stay test split; outcomes by AUROC/AUPRC for mortality and 10-class accuracy for LOS, with stay-level bootstrap CIs (1000 resamples).

Against three EHR foundation models (in-domain ETHOS and EHRMamba, out-of-domain CLMBR-T-base) and a constant-prior floor, CLINTIC wins every column with non-overlapping CIs (Table 4). The gap to CLMBR-T-base localizes the cost of out-of-domain pretrained representations vs. end-to-end learning of this cohort's dynamics; the residual gap to ETHOS/EHRMamba isolates treatment-conditioned latent SDE dynamics over autoregressive token prediction. We note that the EHR-FM baselines are pretrained on broader corpora and then probed on this task, while CLINTIC is trained end-to-end on the target cohort; the comparison thus tests one-shot end-to-end training of a continuous-time treatment-conditioned trajectory model against the prevailing pretrain-tokens-then-probe paradigm at matched downstream evaluation, with apples-to-apples pretraining-corpus matching deferred to follow-up. On compute, CLINTIC's MIMIC training is end-to-end at 10^{15} FLOPs (one H100-day, bfloat16); ETHOS, EHRMamba, and CLMBR-T-base each consume substantially larger pretraining footprints reported by their authors, so Table 4's gap is achieved at lower end-to-end compute, not more.

4 RELATED WORK

EHR foundation models and clinical prediction. Most predictive work on cancer EHRs frames outcomes as static classification at fixed times (Rajkomar et al., 2018; Harutyunyan et al., 2017;

Tomašev et al., 2019; Chen et al., 2025; Adam et al., 2026). Transformer EHR foundation models (BEHRT (Li et al., 2020b), Med-BERT (Rasmy et al., 2021), ETHOS (Renc et al., 2024), Med-Tok (Su et al., 2025)) improve visit-level next-event prediction from structured codes, typically pretrained on MIMIC-IV (Johnson et al., 2023). CLINTIC is generative and continuous-time: we roll out a 26×5 trajectory grid under a chosen treatment and downstream tasks become probes on that grid.

Neural SDEs and continuous-time clinical dynamics. Li et al. (2020a); Kidger et al. (2021) developed scalable latent SDEs in the broader Neural ODE/SDE family (Chen et al., 2018; Oksendal, 2013). Latent ODEs (Rubanova et al., 2019) and GRU-ODE-Bayes (Brouwer et al., 2019) model irregular clinical series; Durso-Finley et al. (2024) applied neural SDEs to multiple sclerosis. Closest to our setup, Pakravan et al. (2024) introduce a pharmacology-informed neural SDE for treatment-effect identification on stochastic health signals, Zhang et al. (2024) train clinical neural SDEs simulation-free via flow matching, and Aslanimoghanloo et al. (2025) fit a latent SDE on simulated PKPD/ICU data with explicit ITE estimation. Our contribution against this line is heterogeneous multi-modal EHR ingestion at scale and treatment conditioning via an MDP-aligned action token that participates in attention but is excluded from the state update, preventing same-week leakage and yielding action-autoregressive dynamics that admit the leakage-free landmark KM and regimen-substitution probe in §3.6.

World models and counterfactual outcome estimation. The world-model paradigm (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2023) learns a latent simulator one can roll out in imagination; Xu et al. (2026) brought it to sepsis decision support, and concurrent work (Ding et al., 2025; Mu et al., 2026) extends the idea to imaging-driven oncology and EHR-based hospitalization simulation. We adopt dream-then-decode with a continuous-time SDE in place of discrete latents, regularized by JEPA (LeCun et al., 2022; Assran et al., 2023). For counterfactual outcome estimation: CRN (Bica et al., 2020) balances treatment histories adversarially, G-Net (Li et al., 2021) runs neural g-computation, TE-CDE (Seedat et al., 2022) uses controlled differential equations, and representation-balancing methods (Shalit et al., 2017) bound ITE. These typically take a small fixed feature vector; CLINTIC provides the multi-modal generative substrate that admits action-conditioned counterfactual rollouts by construction, with end-to-end counterfactual evaluation deferred to follow-up.

5 CONCLUSION

We introduced CLINTIC, a treatment-conditioned latent neural SDE for modeling patient clinical trajectories. The same recipe attains state-of-the-art accuracy and calibration on two distinct real-world cohorts, the **MSK CCDE** oncology release and the **MIMIC-IV** ICU benchmark; its latent dynamics alone out-stratify every individual NSCLC biomarker on lung-subset 52-week OS, and a leakage-free, observational counterfactual regimen-substitution probe read off the same dynamics gives a comparable stratifier on the same cohort, all without a dedicated survival or counterfactual head. An IsoFLOP sweep spanning four orders of magnitude in compute fits a Chinchilla-form frontier whose size exponent on this cohort is $N^*(C) \propto C^{0.25}$, roughly half the autoregressive language-model value and well below the ~ 0.5–0.6 exponents reported for tokenized next-event EHR transformers (Zhang et al., 2025; Waxler et al., 2025); cohort, objective, and tokenization differ across these reports, so the gap is best read as locating this measurement in a more data-bound regime rather than as a property of the SDE backbone in isolation. The significance is twofold: methodologically, action-conditioned stochastic latents give a single backbone that handles classification, calibration, survival-style stratification, and counterfactual probing without task-specific heads; strategically, on this cohort the measurement points to cohort breadth, not architectural novelty, as the binding constraint on further scale-up, and CLINTIC's event-vector interface is designed precisely to make the multi-institutional aggregation that this lever identifies as the empirical follow-up a tractable next step.

REFERENCES

Irsyad Adam, Zekai Chen, David Laprade, Shaun Porwal, David Laub, Erik Reinertsen, Arda Pekis, and Kevin Brown. The patient is not a moving document: A world model training paradigm for

longitudinal ehr. *ArXiv*, abs/2601.22128, 2026.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Muhammad Aslanimoghanloo, Ahmed Elgazzar, and Marcel van Gerven. Generative modeling of clinical time series via latent stochastic differential equations. *ArXiv*, abs/2511.16427, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

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

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *Neural Information Processing Systems*, 2019.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Kristjanson Duvenaud. Neural ordinary differential equations. *Neural Information Processing Systems*, 2018.

Zekai Chen, Arda Pekis, and Kevin Brown. Building the ehr foundation model via next event prediction. *ArXiv*, abs/2509.25591, 2025.

Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. Clarity: Medical world model for guiding treatment decisions by modeling context-aware disease trajectories in latent space. *ArXiv*, abs/2512.08029, 2025.

Joshua Durso-Finley, Berardino Barile, Jean-Pierre Falet, Douglas L Arnold, Nick Pawlowski, and Tal Arbel. Probabilistic temporal prediction of continuous disease trajectories and treatment effects using neural sdes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 400–410. Springer, 2024.

Elizabeth A. Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H. Schwartz, Daniel J. Sargent, Robert Ford, Janet Ellen Dancey, Susan G. Arbuck, S. Gwyther, Margaret Mooney, Lawrence Rubinstein, Lalitha K Shankar, Lori E. Dodd, Robert M Kaplan, Denis Lacombe, and Jaap Verweij. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45 2:228–47, 2009.

Adibvafa Fallahpour, Mahshid Alinoori, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *ArXiv*, abs/2405.14567, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

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

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and A. G. Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 30016–30030, 2022.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

David A. Karnofsky. The clinical evaluation of chemotherapeutic agents in cancer. 1949.

Patrick Kidger, James Foster, Xuechen Li, and Terry J Lyons. Neural sdes as infinite-dimensional gans. In *International conference on machine learning*, pp. 5453–5463. PMLR, 2021.

Yann LeCun et al. A path towards autonomous machine intelligence. *Open Review*, 62(1):1–62, 2022.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, et al. G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Machine Learning for Health*, pp. 282–299. PMLR, 2021.

Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David K Duvenaud. Scalable gradients and variational inference for stochastic differential equations. In *Symposium on Advances in Approximate Bayesian Inference*, pp. 1–28. PMLR, 2020a.

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

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. *International Conference on Machine Learning*, 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*, abs/2602.03569, 2026.

M. M. Oken, Richard H. Creech, Douglass C. Tormey, John Horton, Thomas E. Davis, Eleanor McFadden, and Paul P. Carbone. Toxicity and response criteria of the eastern cooperative oncology group. *AMERICAN JOURNAL OF CLINICAL ONCOLOGY*, 5:649–656, 1982.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Samira Pakravan, Nikolaos Evangelou, Maxime Usdin, Logan Brooks, and James Lu. From noise to signal: Unveiling treatment effects from digital health data through pharmacology-informed neural-sde. *ArXiv*, abs/2403.03274, 2024.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

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

Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7(1):256, 2024.

Yulia Rubanova, Tian Qi Chen, and David Kristjanson Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Neural Information Processing Systems*, 2019.

Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, and Mihaela van der Schaar. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. *arXiv preprint arXiv:2206.08311*, 2022.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Ethan H. Steinberg, Kenneth Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam Haresh Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, pp. 103637, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Darabi Sahneh, and Marinka Zitnik. Multimodal medical code tokenizer. *ArXiv*, abs/2502.04397, 2025.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shane Waxler, Paul Blazek, D. White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, H Voeller, Karen K. Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew J. Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale. *ArXiv*, abs/2508.12104, 2025.

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.

Sheng Zhang, Qin Liu, Naoto Usuyama, Cliff Wong, Tristan Naumann, and Hoifung Poon. Exploring scaling laws for ehr foundation models. *ArXiv*, abs/2505.22964, 2025.

Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis L. Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modeling. *ArXiv*, abs/2410.21154, 2024.

AI USE STATEMENT

We used ChatGPT to assist with manuscript organization and language editing, provide critical feedback on the experimental methodology and statistical reporting, and refine the interpretation and presentation of results. Separately, the Claude models described in Section 3.3 were evaluated as experimental baselines.

ETHICS STATEMENT

This study analyzes retrospective clinical records from MSK CCDE and MIMIC-IV. [State the applicable institutional review board approval, exemption, or determination that additional review was not required, together with the relevant data-access authorization and consent or consent-waiver provisions.] Clinical records contain sensitive information, and their processing and any use with external model services must comply with the applicable data-use agreements. [Describe the privacy safeguards actually applied, including the authorization and safeguards governing patient-record inputs to the Claude baselines.]

ClinTic is a research model and is not intended to guide patient care without further validation. Observational treatment assignment, incomplete documentation, and differences in patient populations and clinical practice may introduce confounding and systematic prediction errors. Performance may vary across demographic and clinical subgroups; aggregate results do not establish equitable performance. Action-substitution analyses do not identify causal treatment effects and should not be interpreted as treatment recommendations. Clinical deployment would require external and prospective validation, subgroup assessment, and appropriate clinical oversight.

REPRODUCIBILITY STATEMENT

The model architecture, treatment-conditioned dynamics, and training objectives are described in Section 2 and Appendix D. Cohort definitions, modality registries, and event preprocessing are documented in Section 3.1 and Appendices B and C. Evaluation procedures and metrics are described in Section 3.2 and Appendix E, with baseline configurations in Appendix G. Appendix H provides the per-configuration scaling results, and Appendix J reports the probabilistic evaluation. Reproduction using the underlying clinical records is subject to the respective datasets' access requirements.

A APPENDIX

B COHORT INSTANTIATIONS

Table 5: Cohort instantiations of the same CLINTIC framework. Modality registry, observation targets, and step cadence are configured per cohort; the encoder, SDE, decoder backbone, JEPA branch, and training recipe are shared.

	MSK CCDE (oncology)	**MIMIC-IV v3.1 (ICU)**
Patients / units	23,320 patients	68,546 ICU stays
Train / test split	18,656 / 4,664	54,825 / 13,721
Events / unit	2.38M total ($\sim$102/pt)	42.8M total ($\sim$625/stay, max 1024)
Step cadence	weekly, 26-step dream	hourly, 24-step dream
Encode window	up to years pre-dream	first 24 h of ICU stay
Modalities (#)	6: assessment, imaging, pathology, treatment, genomics, context	5: vitals, labs, neuro/output, treatment, context
Treatment events	meds, regimens, surgery, radiation	vasopressors, inotropes, sedatives, antimicrobials, fluids, blood, vent., lines
Observation targets	5 MD-authored categoricals: ECOG, KPS, DSt, DAct, DCh	68 cont. channels (Smooth L1); stay-level mortality + 10-class LOS

Table 6: Cohort population statistics. MSK CCDE counts are reported on the lung sub-cohort used for the data report; MIMIC-IV counts are over the full first-admission ICU cohort (adults, $\geq$ 24 h stays).

MSK CCDE (lung sub-cohort)	
Statistic	**Value**
Patients	6,115
Train / test (80/20)	4,892 / 1,223
Total event rows	555,892
Events/pt (med., p95)	74, 220
Source categories	31 (in 6 modalities)
Time encoding	days-since-birth
Median age at last event	70.7 y
Any treatment	91.2%
Sys. / surg. / rad.	79.7 / 70.1 / 19.3%
Genomic data	37.5%
Death date present	44.5%
Tokens/pt (median)	$\sim$6,600

MIMIC-IV v3.1 (ICU)	
Statistic	**Value**
ICU stays	68,546
Train / test (80/20)	54,825 / 13,721
Total event rows	$\sim$42 M
Events/stay (med.)	$\sim$625 (cap 1024)
Event types	135 (68 obs + 33 act.)
Time encoding	hours-since-admission
Observation window	fixed 48 h
Encode / dream split	24 h / 24 h
Median LOS	1.9 d
In-hospital mortality	11.1%
Static features	5
Tokens/event	$\leq$12

Table 7: Per-cohort modality registry and decoded targets. Modalities are pluggable sub-encoders behind the shared d_e-dim event-vector interface (§2.1); targets are read out per dream step (or, for stay-level outcomes, from the final latent).

MSK CCDE – oncology, weekly (W=26)

Modality	**Source categories**
`assessment`	assessment, diagnosis, panel, comorb.
`imaging`	imaging, cancer site, scan site
`pathology`	pathology, specimen, PDL, MMR, MSI
`treatment`	medication, regimen, surgery, radiation
`genomics`	mutation, SV, CNA, copy-number segment
`context`	demographics, family history
Target (K)	**Semantics**
ECOG (6)	performance status, 0 active → 5 dead (ord.)
KPS (11)	Karnofsky 0–100 step 10 (ord.)
DiseaseState (3)	localized / recurrent / distant (ord.)
DiseaseActivity (3)	no / unknown / yes (nom.)
DiseaseChange (5)	RECIST: remission → progressing (ord.)

MIMIC-IV v3.1 ICU – acute, hourly (W=24)

Modality	**Source categories**
`vitals`	HR, BP (S/D/MAP), SpO_2, RR, temperature, ventilator
`labs`	chemistry, hematology, blood gas, coagulation
`neuro/output`	GCS (eye/verbal/motor), urine, chest-tube
`treatment`	vasopressors, sedatives, antimicrobials, fluids, ventilation
`context`	demographics, admission type, careunit
Target	**Semantics**
68 channels (cont.)	per-hour Smooth-L1 reconstruction
mortality (2)	in-hospital, stay-level
LOS (10)	length-of-stay decile, stay-level

C Data vocabulary and event serialization

Gene vocabulary. The genomics modality emits a dense 541×6 matrix per genomics event. Rows index a fixed gene vocabulary built from the union of MSK-IMPACT panel versions in CCDE; we keep all genes that appear with at least one alteration in the train split (after filtering out germline and benign annotations), yielding 541 genes. The 6 channels are: **missense** (count of missense substitutions), **truncating** (nonsense, frameshift, splice), **in-frame** (in-frame insertions/deletions and in-frame mutations not classified as missense), **CNA** (signed copy-number alteration $\in\{-2,-1,0,1,2\}$ for deep deletion through high-level amplification), **SV** (count of structural variants and fusions touching the gene), and **VAF** (max variant-allele frequency across reported variants on the gene, $\in[0,1]$). Channels are L1-normalized within event before passing through the $3246\to512\to d_e$ MLP.

Non-genomics serialization. Each non-genomics event is rendered as a 1–3 line text block: a header naming the source category (e.g. `[imaging]`) followed by key/value rows from the original schema, joined with newlines. Absolute dates, time-of-death, and last-known-alive fields are stripped before serialization; relative timing is delivered separately via the continuous-time RoPE (§2.1). Free-text fields (impressions, narrative, oncology notes) are pre-truncated to fit the 512-subword ClinicalBERT window after concatenation with the structured rows.

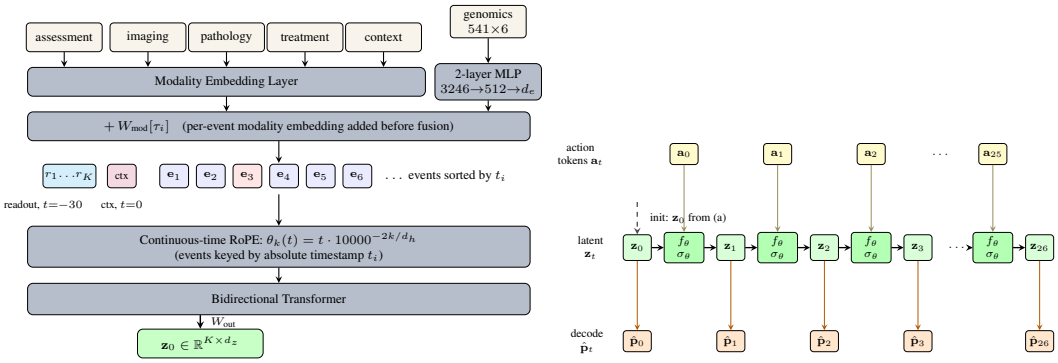
Event grouping. Events sharing a (patient, day, modality) tuple are merged into a single parent event whose serialization concatenates child rows; this prevents the same imaging study from appearing as two events when imaging-cancer-site and imaging-scan-site are recorded separately. The resulting per-patient sequence is truncated to the most recent 256 events relative to dream_start when longer.

D Detailed architecture

Figure 5 expands Figure 1: the upper panel shows per-modality fusion inside the event encoder; the lower panel shows the action–latent attention pattern inside one Euler–Maruyama step of the SDE rollout.

E Per-observation metrics

Table 8 gives the full per-observation metric definitions. We choose threshold AUROC for ordinal targets at the clinically meaningful boundary rather than overall multi-class AUROC: ranking is what survives miscalibration, and ranking against a single threshold ranks against the decision the metric is meant to drive. For DiseaseActivity, which is nominal (Yes vs. No vs. a clinically distinct



(a) Multi-modal event encoder. (b) Latent SDE rollout.

Figure 5: **Detailed architecture, complementary to Figure 1. (a) Multi-modal event encoder.** Each clinical event enters through a modality-specific sub-encoder: five textual modalities (assessment, imaging, pathology, treatment, context) share a single fine-tuned ClinicalBERT (mean-pool over $\leq$ 512 subword tokens, then a linear projection to d_e); genomics flattens its $541{\times}6$ matrix through a 2-layer MLP. A learned modality embedding $W_{\text{mod}}[\tau_i]$ is added so the downstream Transformer can route by channel without losing the shared event-embedding space. Events are sorted by absolute timestamp and packed (variable-length, no padding) with $K{=}16$ learnable readout tokens at $t{=}-30$ d and the demographics anchor at $t{=}0$ prepended; continuous-time RoPE keys every position by its absolute timestamp t_i. A 6-layer bidirectional Transformer with FlashAttention-3 emits $\mathbf{z}_0 \in \mathbb{R}^{K \times d_z}$ from the readout positions only. **(b) Treatment-conditioned latent SDE.** The K latent tokens are evolved 26 weekly bins. At step t the drift Transformer attends *jointly* over $[\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(K)}, \mathbf{a}_t]$ and returns K updated tokens; the action $\mathbf{a}_t$ (mean-pool of treatment-event embeddings in $[t, t{+}1]$, or a learned no-treatment placeholder) participates as conditioning, while the SDE update is over the K latent slots only. The state-dependent diffusion $\sigma_\theta(\mathbf{z}_t)$ widens variance where future trajectories are uncertain. Euler-Maruyama integrates with $S{=}4$ sub-steps per week. **MDP alignment:** the per-week observation decoder reads $\mathbf{z}_t$ *before* action $\mathbf{a}_t$ enters the drift, making the dynamics autoregressive in treatment and immune to same-week leakage; this property underlies the landmark construction in §3.6. The JEPA branch (cf. Figure 1) is omitted here for clarity.

Unknown class), macro-F1 weights the three classes equally despite the strong Yes-prevalence in the test split. ECOG and KPS use quadratic-weighted Cohen's κ, the standard for ordinal agreement.

Table 8: Primary per-observation metric and the clinical decision it ranks against.

Observation	**Primary metric**	**Decision being ranked**
ECOG	QWK (ordinal κ)	Performance-status ordinal agreement
KPS	QWK (ordinal κ)	Performance-status ordinal agreement
DiseaseState	AUROC at $y \geq 2$	Metastatic vs. non-metastatic
DiseaseActivity	Macro-F1 over 3 classes	Active disease (Yes / Unknown / No)
DiseaseChange	AUROC at $y \geq 4$	RECIST-PD-analogue progression

We report patient-level bootstrap CIs (200 resamples over patients) on every primary metric. Resampling at the patient level rather than the cell level preserves within-patient trajectory correlation and produces conservative intervals; in early experiments cell-level bootstrap intervals were $\sim 2\times$ tighter and consistently overstated significance.

F Dream-window qualitative analyses

Figure 6 shows the latent dynamics for the same patient as Figure 2: per-week L2 drift impulses concentrate at treatment weeks rather than co-occurring uniformly, evidence that the action conditioning carries the intra-trajectory variation rather than being absorbed into a smooth flow. Figure 7 reports per-observation primary metrics across the full 26-week horizon on the held-out test split: metrics are approximately flat across the horizon for the high-data observations (ECOG, KPS, DiseaseState, DiseaseActivity), with no long-horizon collapse at the SDE rollout's far end; DiseaseChange has wider CIs at later weeks reflecting the lower per-week observation density of RECIST events.

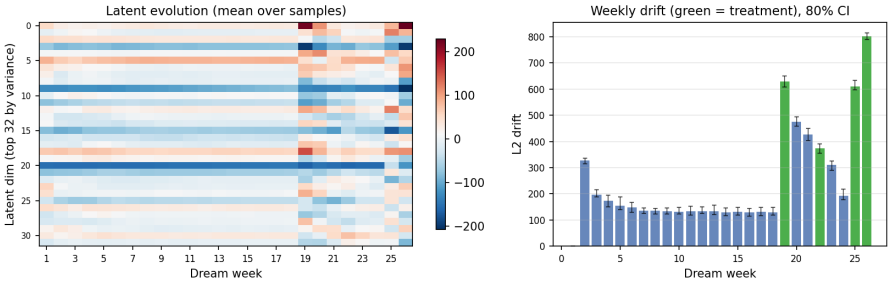


Figure 6: Latent dynamics for the patient in Figure 2. Top-32 latent dims (left) and per-week L2 drift with 80% CI (right; green = treatment weeks). Drift impulses concentrate at treatment weeks.

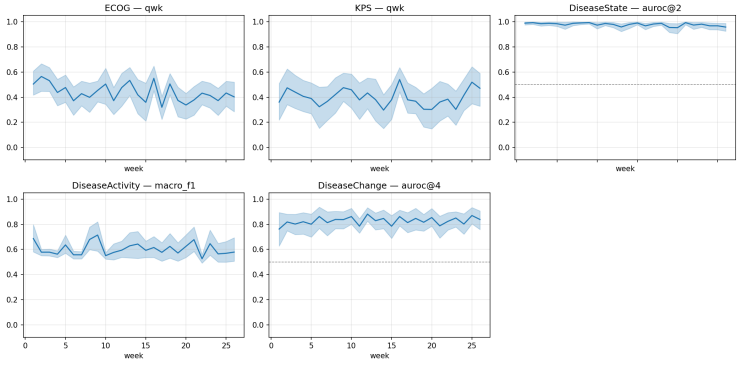


Figure 7: Primary metric per observation across the 26-week horizon with bootstrap CIs on the 4,664-patient test split.

G BASELINE CONFIGURATIONS

Trained baselines. All trained baselines use the same train/test patient split as CLINTIC and the same dense 26×5 evaluation grid, so per-cell metrics are directly comparable. The per-cell tabular baselines (LogReg, XGBoost) build a per-patient feature vector from the encode window (counts and one-hot indicators over event categories, age, sex, top-50 most-altered genes) and emit one independent prediction per (patient, week, obs) cell. The recurrent counterfactual baselines (CRN, Causal Transformer, Latent ODE, Dynamic-DeepHit, G-Net) use their authors' reference implementations adapted to the CCDE schema; treatment indicators enter as the standard binary action vector. Latent ODE has no treatment input and serves as a pure continuous-time dynamics ablation. G-Net uses Monte-Carlo G-computation rollouts at evaluation time. We tuned learning rate per baseline by val loss; other hyperparameters follow the original publications.

Zero-shot LLMs. **Claude** models see each test patient's pre-prediction event log reformatted as a chronological text record. Absolute dates are stripped; events carry only relative day offsets from the patient's first event. A compact demographics line and a top-altered-gene summary are pinned at the start of the prompt as always-on context. The model is asked to emit, for every (week, obs) cell in the dream window, a class-probability distribution over the appropriate label set, in a strict JSON schema (one object per week, observations as keys, class names as keys, probabilities as values summing to 1). The system prompt instructs the model to attend to the per-event temporal structure, to treat treatment events as conditioning rather than outcomes, and to avoid degenerate uniform distributions when the pre-window evidence supports a confident prediction.

Parsing pipeline. The LLM completion is parsed with a permissive JSON extractor that tolerates trailing commentary, markdown code fences, and partial outputs. Missing weeks default to a uniform distribution over the obs's class set; missing classes within an emitted week default to zero before re-normalization. Probabilities are clipped to $[10^{-6}, 1-10^{-6}]$ before metric computation. We retry once on hard parse failures; the retry rate was $< 1\%$ on both LLM baselines.

Sample-efficiency note. The two LLMs receive zero supervised cells from the train split, while every other method in 1 sees the full $\sim$18.7k-patient train set. The comparison is therefore deliberately asymmetric; the result of interest is not which side wins on average but which observation classes break the LLM (dynamic ordinals: ECOG, KPS, DiseaseChange) versus which it nearly closes (structural extent: DiseaseState).

H SCALING-LAW ANALYSIS AND PER-CELL ISOFLOP RESULTS

This appendix backs the compressed analysis in §3.4: we expand on the operating-point reading of the fit, the downstream-metric behavior on and off the envelope, and the methodological caveats; the per-cell sweep that drives all three is in Table 9.

Operating point and remaining headroom. The fit also locates the operating point on the curve. At $C = 10^{15}$ FLOPs the predicted L^* is 0.651 and the observed depth-6 cell attains 0.6552, putting the production model within $\sim$0.009 of the irreducible-loss asymptote $E = 0.646$. Two readings of E are useful. As a fitted quantity it bounds how compressible weekly oncology trajectories are under the current modality registry and supervision targets; reductions below it require richer covariates (volumetric imaging, structured pathology, physiologic waveforms) or denser per-step supervision rather than more parameters. As an operating budget, halving the excess loss above E costs $\sim3\times$ compute under the fit, but only along the D-axis: with cohort fixed at 23,320 patients the gain asymptotes; with cohort growth the law projects continued recovery. The exponent locates the bottleneck precisely, the asymptote sets the price.

Downstream tracks the envelope, then collapses outside it. The full per-cell results in Table 9 confirm that the val-recon ordering on the envelope matches the downstream-metric ordering, but only along the envelope. At $C{=}3{\times}10^{12}$, depths ≥ 6 all post-collapse to majority-class behaviour (DChg AUROC@4 $\approx$0.5, DAct F1 $\approx$0.27); at $C{=}10^{15}$ the same depths recover and the $L{=}6$ sweep cell attains the global maxima on both DState (0.982) and DChg (0.839), in addition to the lowest

val recon (0.6552) across the entire 41-cell sweep. Per-cell rows in Table 9 are independently-seeded IsoFLOP-target runs; the headline production model evaluated in Table 1 is a separate run at this same L=6 configuration, so its downstream values differ marginally from this row (e.g., DChg AUROC@4 0.827 in Table 1 vs. 0.839 here). The recipe and supervision targets are shared across cells, so the sharpness of this transition is driven by the compute budget alone, not by per-cell tuning. The location of the post-collapse cliff is itself a useful operational signal: capacity that outruns its data does not gracefully degrade in this regime, it falls off a constant predictor.

Caveats on the fit. Three are worth flagging. (i) The envelope is taken over depths ≤ 8, the canonical Chinchilla window where the per-budget min is well-defined; depths 12 and 16 sit above the envelope at every budget but were run to verify the optimal-depth schedule eventually stalls rather than continuing to climb. (ii) Both the fit and the data-bound diagnosis are established on a single cohort over 3×10^{12} to 10^{15} FLOPs; whether the same exponents hold under multi-institutional aggregation — the regime the law itself motivates — remains to be measured, and is the most important follow-up. (iii) The minimized loss is the trajectory-modeling objective (per-step reconstruction with JEPA auxiliary), not a downstream task loss; the downstream-metric agreement on the envelope is encouraging but not part of the fit, and the same exponent need not transfer to e.g. a survival or treatment-effect objective on the same data.

Per-cell sweep. Table 9 reports primary downstream metrics for every cell in the IsoFLOP sweep, in addition to the val-recon objective used to fit Figure 3. Cells are grouped by FLOP budget and sorted by depth within each group. Primary metrics match the main-text convention: ECOG and KPS use quadratic-weighted κ, DiseaseState uses AUROC at the metastatic boundary ($y \geq 2$), DiseaseActivity uses macro-F1, and DiseaseChange uses AUROC at the RECIST-PD boundary ($y \geq 4$). The cell at the production configuration (L=6, F=10^{15}) is shaded.

I Additional Kaplan–Meier analyses

Per-regimen landmark KM panels. Figure 8 shows the four largest named-regimen cohorts in the test split, each at the landmark week reported in Table 3. The Nivolumab panel (T=13) shows the strongest separation in the paper (HR 6.35 [3.06, 13.17], log-rank $p = 7.1\times10^{-7}$), with the high- and low-s_T arms diverging immediately at the landmark and remaining well separated through 52 weeks. Pembrolizumab (T=4) and FOLFIRINOX (T=4) show similar early divergence at smaller magnitude; the Carboplatin/Paclitaxel arm has the highest HR point estimate among non-immunotherapy regimens but with a wide CI driven by only 12 events. Across regimens the landmark score discriminates outcome trajectories beyond the landmark itself, consistent with the action-autoregressive design.

Caveats. The Carboplatin/Paclitaxel HR point estimate is high but driven by few events; we report it for completeness. Cohort sizes for ALK and MET in the NSCLC test subset are small (single-digit alteration counts), and the corresponding biomarker HRs in Figure 4 (top) should be read accordingly.

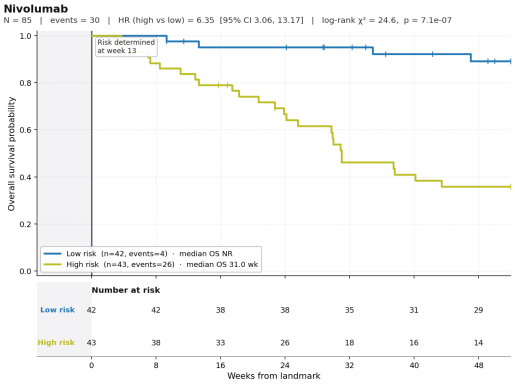
J Calibration of per-cell predictions: full SDE vs −SDE

The component ablation in §3.5 reports point-estimate metrics; calibration is reported separately here so the substantive comparison is read against a proper scoring rule rather than against accuracy alone. We compare the full SDE rollout against the −SDE ablation ($\sigma\equiv0$, every other design choice held fixed) on identical (patient, week, obs) cells of the held-out MSK CCDE test split, using two metrics: the multi-class Brier score (a strictly proper scoring rule) and Expected Calibration Error (ECE, 15 bins, sharpness-agnostic). Lower is better for both. Confidence intervals are patient-level bootstrap with 200 resamples.

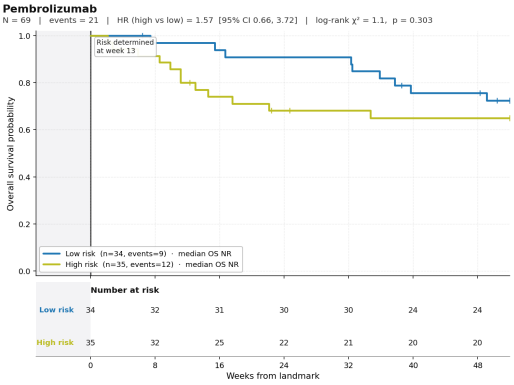
Brier favors the full SDE on every observation, with non-trivial gaps on the two structural targets (DiseaseState −0.021, DiseaseActivity −0.013) where point-estimate metrics in the main-text ablation were already near saturation; on the dynamic targets the gap is smaller but still in the same direction. ECE is mixed at margins below the bootstrap CI half-width on every row, and reads as

Table 9: Per-cell IsoFLOP results across 7 depths × 6 FLOP budgets (41 cells; L=16 at F=10^{15} omitted). Columns: depth L, scaling parameters N, validation reconstruction loss, and the primary metric per observation (higher is better for all five). The cell at the production configuration (L=6, F=10^{15}) is shaded; per-cell rows are independently-seeded IsoFLOP-target runs and so differ marginally on downstream metrics from the headline production run at the same L=6 configuration in Table 1.

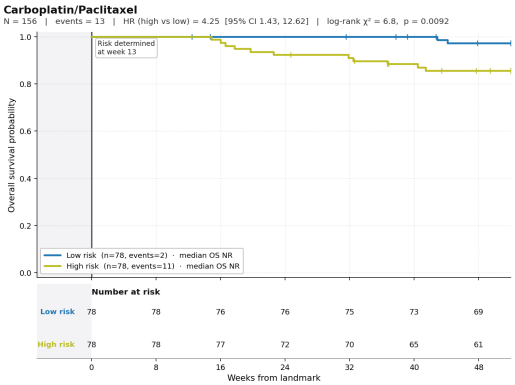
L	N (M)	val recon@	ECOG QWK	KPS QWK	DState AUROC@2	DAct F1	DChg AUROC@4
$F = 3\times10^{12}$ *FLOPs*							
2	1.68	0.7878	0.230	0.156	0.944	0.473	0.706
3	2.50	0.8027	0.229	0.182	0.934	0.487	0.700
4	3.51	0.8548	0.188	0.169	0.911	0.473	0.667
6	8.16	0.9877	0.000	0.009	0.725	0.271	0.563
8	15.09	0.9961	0.000	0.000	0.614	0.271	0.518
12	39.20	1.3594	0.000	0.000	0.702	0.165	0.543
16	79.07	5.6717	0.000	0.004	0.660	0.271	0.501
$F = 1\times10^{13}$ *FLOPs*							
2	1.68	0.7201	0.309	0.279	0.970	0.507	0.743
3	2.50	0.7396	0.245	0.227	0.965	0.510	0.738
4	3.51	0.7464	0.230	0.196	0.965	0.500	0.742
6	8.16	0.8177	0.201	0.105	0.926	0.484	0.689
8	15.09	0.9693	0.098	0.000	0.746	0.271	0.606
12	39.20	1.0049	0.000	0.002	0.771	0.271	0.622
16	79.07	1.7736	0.000	−0.001	0.653	0.271	0.503
$F = 3\times10^{13}$ *FLOPs*							
2	1.68	0.7222	0.351	0.304	0.961	0.500	0.725
3	2.50	0.6874	0.388	0.362	0.968	0.532	0.766
4	3.51	0.6951	0.370	0.350	0.970	0.535	0.749
6	8.16	0.7525	0.266	0.179	0.964	0.502	0.740
8	15.09	0.7733	0.238	0.177	0.947	0.483	0.709
12	39.20	1.0571	0.106	0.056	0.756	0.374	0.641
16	79.07	1.0441	0.000	0.002	0.644	0.271	0.560
$F = 1\times10^{14}$ *FLOPs*							
2	1.68	0.6781	0.408	0.368	0.970	0.578	0.786
3	2.50	0.6579	0.431	0.382	0.973	0.597	0.811
4	3.51	0.6647	0.423	0.358	0.972	0.553	0.798
6	8.16	0.6906	0.381	0.346	0.969	0.539	0.752
8	15.09	0.7309	0.271	0.227	0.964	0.520	0.741
12	39.20	0.7898	0.210	0.177	0.938	0.480	0.703
16	79.07	0.9475	0.106	0.089	0.771	0.350	0.645
$F = 3\times10^{14}$ *FLOPs*							
2	1.68	0.6616	0.401	0.359	0.980	0.613	0.819
3	2.50	0.6710	0.418	0.349	0.976	0.589	0.784
4	3.51	0.6570	0.422	0.391	0.975	0.598	0.827
6	8.16	0.6703	0.382	0.352	0.971	0.580	0.802
8	15.09	0.6791	0.427	0.376	0.969	0.547	0.770
12	39.20	0.7494	0.210	0.264	0.961	0.503	0.733
16	79.07	0.7715	0.236	0.205	0.951	0.488	0.708
$F = 1\times10^{15}$ *FLOPs*							
2	1.68	0.6758	0.411	0.358	0.971	0.563	0.774
3	2.50	0.6673	0.413	0.363	0.978	0.542	0.795
4	3.51	0.6678	0.393	0.358	0.973	0.583	0.804
6	8.16	**0.6552**	**0.428**	0.377	**0.982**	**0.599**	**0.839**
8	15.09	0.6654	0.421	**0.382**	0.969	0.597	0.824
12	39.20	0.7029	0.423	0.357	0.969	0.535	0.750



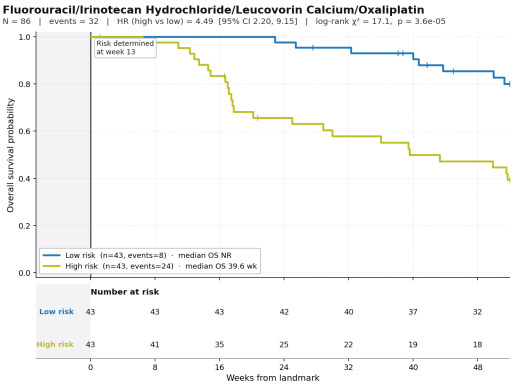
(a) Nivolumab cohort, T=13. HR 6.35 [3.06, 13.17], $p = 7.1\times10^{-7}$, n=85, 30 events.



(b) Pembrolizumab cohort, T=4. HR 2.40 [1.01, 5.67], n=70, 21 events.



(c) Carboplatin/Paclitaxel cohort, T=4. HR 5.60 [1.80, 17.38], n=156, 12 events.



(d) FOLFIRINOX cohort, T=4. HR 2.27 [1.03, 5.02], n=86, 25 events.

Figure 8: Per-regimen landmark Kaplan–Meier curves on the median split of s_T. Each panel shows the high-s_T (red) versus low-s_T (blue) arms beyond the landmark week with 95% confidence bands; events are deaths within 52 weeks of the landmark, censored otherwise.

Table 10: Per-observation calibration on the dense 26×5 test grid: full CLINTIC vs. −SDE ablation. Each cell: mean [95% CI]. **Bold** = better in column-pair; ties broken by point estimate.

	Brier ↓		ECE ↓	
Observation	**CLINTIC**	**−SDE**	**CLINTIC**	**−SDE**
ECOG	**.451 [.434,.469]**	.458 [.438,.477]	.089 [.080,.100]	**.088 [.077,.099]**
KPS	**.578 [.560,.595]**	.579 [.563,.598]	.105 [.093,.119]	**.095 [.083,.109]**
DiseaseState	**.127 [.101,.156]**	.148 [.117,.174]	**.057 [.046,.068]**	.063 [.054,.073]
DiseaseActivity	**.231 [.218,.247]**	.244 [.232,.259]	**.053 [.048,.060]**	.059 [.052,.065]
DiseaseChange	**.432 [.417,.452]**	.449 [.434,.467]	.091 [.082,.101]	**.086 [.077,.096]**

small re-bin noise rather than a systematic effect: ECE collapses each bin to its mean confidence and so is less sensitive to refinement than Brier. The right reading is therefore that the diffusion improves probabilistic calibration in the proper-scoring-rule sense without measurably re-binning the per-class confidence histogram. This is the calibration-flavored gap referenced from §3.5: the −SDE ablation isolates stochasticity within the framework (same encoder, same actions, deterministic ODE rollout), so the Brier difference is attributable to the learned diffusion $\sigma_\theta(\mathbf{z}_t)$ rather than to any other architectural ingredient. We do not claim that this resolves the broader question of stochastic-rollout-derived prediction sets, which remains open.