TUMORCAST: A CANCER WORLD MODEL FOR LONGITUDINAL TUMOR FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Anticipating how a tumor will grow and spread could help clinicians adapt therapy in time, yet forecasting its evolution is difficult: tumors exhibit substantial heterogeneity, and their trajectories depend on anatomical location, current disease state, local tissue context, and the patient's clinical history. Existing approaches often predict clinical outcomes directly from observed CT scans or records, or generate future CT scans, leaving the evolution of CT latent states insufficiently modeled. To address this gap, we propose ***TumorCast***, a cancer world model for longitudinal tumor forecasting that predicts CT latent states conditioned on clinical history, interval treatment, and transition interval. Built on the joint-embedding predictive architecture (JEPA), ***TumorCast*** learns reusable predicted CT latent states through a Representation–Dynamics–Readout curriculum, without reconstructing future CT scans. First, longitudinal CT scan pairs adapt a CT encoder and latent predictor through future-state supervision, while clinical text provides conditioning inputs for state transitions. Next, chained training learns successive state transitions through recursive rollout, with future-state supervision propagating backward through the trajectory so that later prediction errors refine earlier state transitions and predicted intermediate states. Finally, a shared clinical reader (Qwen3.5-9B/27B) combines predicted CT latent states with clinical context to forecast lesion diameter, lesion size change, and liver metastasis status. To support evaluation, we curate a multimodal tumor forecasting benchmark comprising 1,344 patients, 3,606 CT scans with corresponding radiology reports, and 11,728 structured clinical entries covering clinical history and interval treatment, yielding 2,400 forecasting examples across liver and pancreas. Experiments show that ***TumorCast*** outperforms rendered-slice vision-language models (VLMs) and text-only baselines, highlighting the success of our multimodal JEPA design and potential clinical effectiveness.

1 INTRODUCTION

Longitudinal computed tomography (CT) scans are central to monitoring cancer progression and treatment response, revealing changes in lesion size and the emergence of metastatic disease on follow-up CT scans (O'connor et al., 2017). Forecasting these changes before the next CT scan, as illustrated in Figure 1, could support more timely and individualized cancer management. However, tumor evolution is highly heterogeneous and patient-specific (Dagogo-Jack & Shaw, 2018). Each CT scan captures the disease state at a particular time point, yet patients with similar imaging findings may follow substantially different trajectories depending on their clinical history and interval treatment. Forecasting tumor evolution therefore requires modeling how CT latent states evolve over time in the context of each patient's clinical history and interval treatment.

Medical world models provide a natural framework for this problem by representing patient states and learning their evolution over time. Existing approaches, however, primarily learn such dynamics from longitudinal electronic health records (EHRs), predicting clinical events or future patient representations (Mu et al., 2026; Wang et al., 2026; Yang et al., 2026). These representations capture important clinical context but do not directly model the spatial changes observed in follow-up CT. Vision-language models (VLMs) can additionally incorporate imaging evidence, but typically predict outcomes directly from the observed CT without explicitly modeling the intervening latent dynamics (Bai et al., 2024; Google, 2026). Generative approaches instead synthesize future CT scans

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.

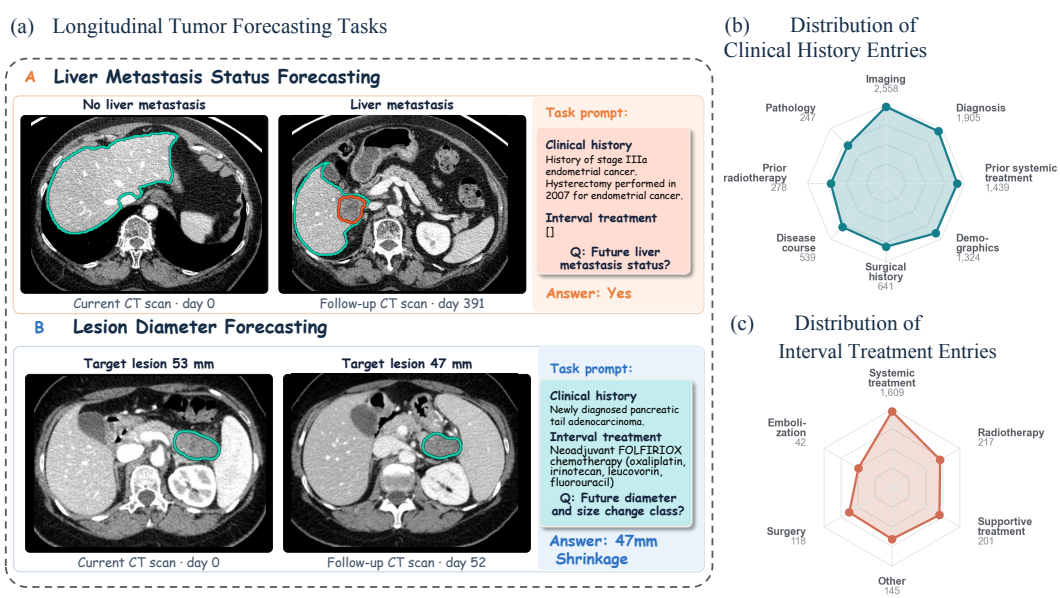


Figure 1: Overview of tumor forecasting tasks and clinical context. (a) Examples of liver metastasis forecasting and lesion size change classification. (b, c) Distributions of structured clinical entries covering clinical history and interval treatment in the dataset.

in pixel space (Mantegna et al., 2026), which is computationally expensive and reconstructs image details that may be unnecessary for tumor forecasting. These advances leave open how to learn latent dynamics conditioned on clinical context across successive transition intervals. Learning these latent dynamics directly from longitudinal CT scans enables follow-up CT scans to supervise CT latent states beyond task-specific labels. The predicted CT latent state then support multiple clinical readouts, motivating tumor forecasting through latent dynamics in a CT latent space.

We propose ***TumorCast***, a cancer world model for tumor forecasting. The model encodes the current organ CT crop into a CT latent state and propagates it with a latent predictor derived from joint-embedding predictive architecture (JEPA) (Assran et al., 2023; 2025). Unlike JEPA models that discard the predictor after representation learning, ***TumorCast*** retains its predictor at inference for state transitions. The latent predictor is explicitly designed to separate conditioning inputs. Clinical history and interval treatment are encoded as text tokens and incorporated through cross-attention, while numerical and categorical conditioning modulates every spatial patch token through AdaLN, using the baseline lesion diameter, transition interval, organ identity, and acquisition phase. This separation prevents numerical and categorical conditioning from competing with a longer clinical text sequence and allows the transition interval to directly control the predicted state transition.

Training the model introduces two coupled challenges: an untrained latent predictor can distort the CT encoder during joint optimization, while a continually changing CT encoder causes the target CT latent space to drift. We address these challenges with a training curriculum that (i) adapts the CT encoder and latent predictor on longitudinal CT scan pairs using an exponential-moving-average (EMA) target encoder; (ii) freezes the CT encoder, aligns the latent predictor to the stabilized CT latent space, synchronizes the target CT encoder, and refines the latent predictor with paired and chained prediction losses. In chained training, the latent predictor consumes its own predicted intermediate state rather than an observed intermediate CT scan, while a follow-up CT scan supervises the endpoint, allowing later errors to propagate through earlier state transitions; (iii) a visual bridge maps the predicted CT latent state into tokens read by a clinical reader, which combines the predicted CT latent state with clinical context to forecast lesion diameter, lesion size change, and liver metastasis status. This design separates dynamics learning from clinical readout while allowing multiple predicted outcomes to be read from a shared predicted CT latent state.

In summary, our contributions are as follows:

• **A cancer world model for tumor evolution.** We formulate tumor forecasting as the prediction of CT latent states, conditioned on clinical context and transition interval. Follow-up CT scans provide

target CT latent states for future-state supervision, enabling tumor forecasting without reconstructing future CT scans.

• **A Representation–Dynamics–Readout curriculum.** We develop a training curriculum that adapts CT latent states, trains state transitions on predicted intermediate states, and connects the predicted CT latent states to a shared clinical reader.

• **A chained training strategy.** We train the latent predictor on its own predicted intermediate states, with follow-up CT scans providing future-state supervision at each step. This improves lesion size change BAcc by 6.4 points for single-step forecasting and 7.0 points for recursive rollout compared with paired training (Table 2).

• **Forecasting performances and insights.** We curate a longitudinal cohort and evaluate single-step forecasting and recursive rollout, with ablations examining the contributions of CT latent state prediction, clinical text conditioning, and chained training to lesion size change classification and liver metastasis forecasting.

2 Related Work

World model architectures. World models span several paradigms (Ha & Schmidhuber, 2018; Hafner et al., 2023; LeCun, 2022). Generative-interactive models such as Genie (Bruce et al., 2024), GAIA-1 (Hu et al., 2023), and GameNGen (Valevski et al., 2025) simulate future observations in pixel space. Latent state-space models such as DreamerV3 (Hafner et al., 2023) instead learn compact dynamics for long-horizon prediction and planning. A complementary line of work focuses on joint-embedding predictive architectures (JEPAs), including I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024), which directly predict target representations in latent space. Subsequent work has extended this paradigm to downstream multimodal dynamics and planning, including V-JEPA 2-AC (Assran et al., 2025), DINO-WM (Zhou et al., 2025) and VL-JEPA (Chen et al., 2026). Motivated by this perspective, ***TumorCast*** represents longitudinal CT scans as a sequence of CT latent states and learns their latent dynamics.

Medical world models and longitudinal forecasting. Medical world models have begun to model longitudinal patient trajectories from EHRs and medical imaging. EHR-based approaches forecast future diagnoses, medications, and clinical events (Kraljevic et al., 2024; Renc et al., 2024; Fallahpour et al., 2024), while MedDreamer (Xu et al., 2026) and Clin-JEPA (Yang et al., 2026) further model latent patient dynamics from longitudinal records. In oncology, imaging-based world models such as MeWM (Yang et al., 2025) synthesize future CT scans for treatment-response modeling and planning, while others model tumor evolution directly in latent space. CLARITY (Ding et al., 2025) couples a LoRA-adapted MRI foundation encoder with an action-conditioned latent transition predictor and survival modeling to evaluate candidate therapies. ***TumorCast*** adopts JEPA-style predictive learning to learn latent dynamics in CT latent space, conditioned on clinical history and interval treatment, and predicts CT latent states that support tumor forecasting.

3D spatial representations for longitudinal modeling. Recent medical VLMs preserve spatial evidence across slices (Lee et al., 2026; Lin et al., 2026; Yi et al., 2026) and learn volumetric anatomical representations (Wu et al., 2025; Hamamci et al., 2026). Some tumor-focused models such as TumorChain (Li et al., 2026) further adopt native 3D CT encoders such as M3D-CLIP Bai et al. (2024) for volumetric tumor analysis, while longitudinal models such as LUMEN (Jiang et al., 2026) reason over observed longitudinal CT scans for prognosis and diagnosis. However, these models analyze observed representations rather than predicted unobserved future states. ***TumorCast*** instead represents CT scans as CT latent states and learns their latent dynamics to obtain predicted CT latent states.

3 Method

We propose ***TumorCast***, a cancer world model that represents the current CT scan as an observed CT latent state, predicts its evolution conditioned on clinical history, interval treatment, and transition interval, and reads predicted outcomes from the predicted CT latent states. The model couples a CT

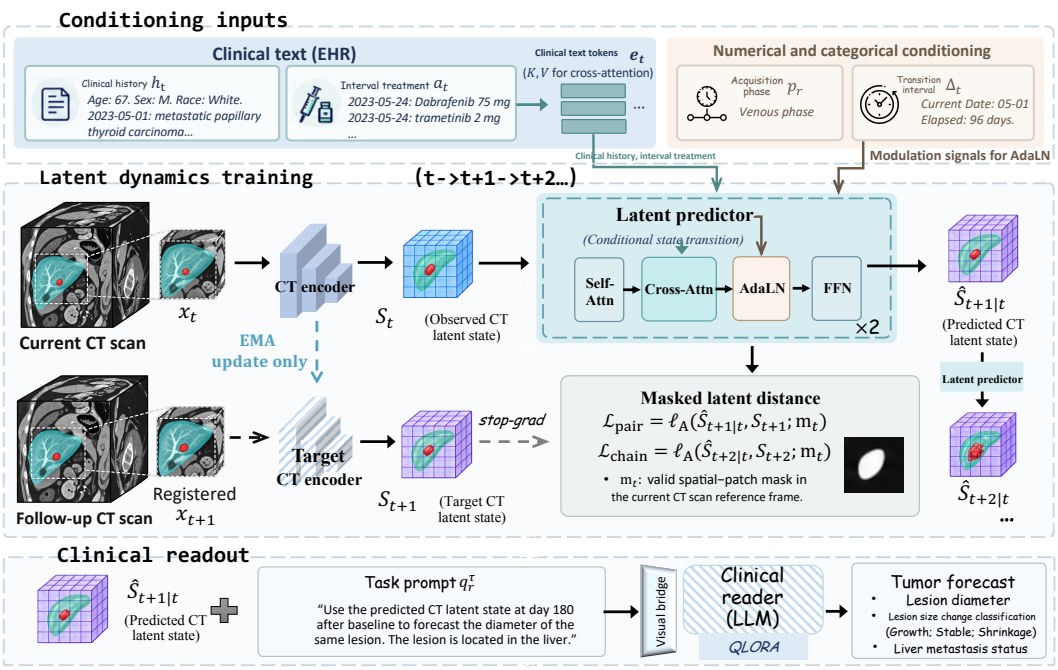


Figure 2: ***TumorCast* Architecture.** The model integrates clinical context with CT latent states, learns state transitions conditioned on interval treatment and transition interval through future-state supervision from follow-up CT scans, and uses a clinical reader for tumor forecasting.

encoder, a latent predictor, and a clinical reader based on a large language model (LLM). Figure 2 presents the overall architecture and training curriculum.

3.1 Cancer World Model Formulation

We define a *cancer world model* as a patient-specific model of tumor evolution with three components: (i) CT encoding that represents the observed organ CT crop as a CT latent state; (ii) latent dynamics that propagate this CT latent state to future time points using clinical context and transition intervals; and (iii) clinical readout that translates predicted CT latent states into predicted outcomes.

Inputs and CT latent state. Consider a patient with observations at times $t_0 < \cdots < t_{K-1}$. At the forecast origin t_0, the available input is $o_0 = (x_0, e_0)$, where x_0 is an organ CT crop and e_0 is clinical text from the patient's EHR, including clinical history and interval treatment a_0. Clinical history includes tumor status, lesion measurements, biochemical indicators, diagnostic information, and treatment history, when available. Interval treatment a_0 describes treatment during the transition interval between t_0 and t_1. The organ CT crop provides spatial evidence of the tumor and surrounding anatomy, while the clinical text provides complementary clinical context. Clinical context comprises clinical history and interval treatment.

We encode the organ CT crop and clinical text separately:

$$s_0 = E_\theta(x_0), \qquad h_0 = T_\omega(e_0), \tag{1}$$

where E_θ is a 3D CT encoder and T_ω is a text encoder that remains frozen. We initialize E_θ from M3D (Bai et al., 2024), retaining its global token and spatial patch tokens from one or more physical-space views, and use Qwen3.5-2B as T_ω. The observed CT latent state s_0 is a spatially structured representation derived solely from the organ CT crop. Clinical text tokens h_0 provide a separate conditioning input to the latent predictor.

Latent dynamics. Starting from the observed CT latent state s_0, the model predicts its evolution conditioned on clinical text tokens and numerical and categorical conditioning inputs. For the transition from t_r to t_{r+1}, let p_r denote acquisition metadata and $\Delta_r = t_{r+1} - t_r$ the transition

interval. The forecast horizon at t_r is the elapsed time from the forecast origin t_0 to t_r. Writing $\xi_r = (p_r, \Delta_r)$, we define

$$\hat{s}_{r+1|0} = F_\psi\big(\hat{s}_{r|0}, h_r, \xi_r\big), \qquad \hat{s}_{0|0} = s_0, \tag{2}$$

where $\hat{s}_{r|0}$ denotes the predicted CT latent state at t_r obtained from the baseline observation.

As shown in Figure 2, the latent predictor F_ψ is a lightweight Transformer with three inputs: the current CT latent state, clinical text tokens, and numerical and categorical conditioning inputs. Each layer of the latent predictor contains submodules of self-attention, cross-attention, AdaLN, and FFN. Self-attention models interactions among CT latent state tokens, while cross-attention uses CT latent state tokens as queries and clinical text tokens as keys and values. Numerical and categorical conditioning inputs, including transition interval and acquisition phase, condition the latent predictor through AdaLN (Peebles & Xie, 2023) modulation. Residual connection was added between the input and the output of each layer. The latent predictor outputs a predicted CT latent state with the same token structure and feature dimensionality as the observed CT latent state, allowing its output to serve as the predicted intermediate state for the next state transition.

Outputs and clinical tasks. The clinical reader decodes the predicted CT latent state $\hat{s}_{r|0}$ under the task prompt q_r to produce a generated answer, from which the predicted outcome is extracted:

$$\hat{u}_r \sim p_{\mathrm{LLM}}(\cdot \mid \hat{s}_{r|0}, q_r), \qquad \hat{y}_r = \mathrm{Extract}(\hat{u}_r), \tag{3}$$

where $\hat{u}_r$ denotes the generated answer and q_r contains the task question and clinical context. The visual bridge maps the predicted CT latent state to tokens for the clinical reader, which is trained through instruction tuning. For lesion diameter forecasting, $\hat{y}_r$ is the predicted lesion diameter. For lesion size change classification, the Shrinkage, Stable, or Growth class is computed from $\hat{y}_r - \hat{y}_{r-1}$ according to the clinical criteria specified in the evaluation protocol, with $\hat{y}_0 = y_0$ denoting the observed baseline diameter. For liver metastasis forecasting, $\hat{y}_r \in \{0, 1\}$ indicates the predicted absence or presence of liver metastasis at follow-up.

3.2 Learning the Cancer World Model

The Representation–Dynamics–Readout curriculum trains ***TumorCast*** through representation adaptation (Stage A), dynamics learning (Stage B), and clinical reader adaptation (Stage C). Target CT latent states encoded from follow-up CT scans provide future-state supervision in Stages A and B; reference answers supervise Stage C. Throughout training, the text encoder T_ω remains frozen.

3.2.1 Stage A: Representation Adaptation

In this stage, we jointly train the CT encoder E_θ and latent predictor F_ψ for tumor forecasting. To focus representation learning on the organ of interest, we construct a spatial mask from a precomputed organ segmentation, expanded outward by a padding margin to include surrounding tissue. The mask is mapped to the patch-token grid.

For each CT scan pair, we define:

$$\begin{aligned} s_r &= E_\theta(x_r), \\ h_r &= T_\omega(e_r), \\ \hat{s}_{r+1|r} &= F_\psi(s_r, h_r, \xi_r), \\ s_{r+1} &= \mathrm{sg}[E_{\bar{\theta}}(\tilde{x}_{r+1})]\,, \end{aligned} \tag{4}$$

where sg denotes the stop-gradient operation and $\tilde{x}_{r+1}$ is the registered organ CT crop from the follow-up CT scan, aligned to the spatial reference of the current organ CT crop x_r. These CT latent states are constructed separately for each pair. The follow-up CT scan provides a target CT latent state for future-state supervision only; follow-up clinical text is not used to construct the target CT latent state. The target CT encoder follows the EMA update $\bar{\theta} \leftarrow \rho\bar{\theta} + (1-\rho)\theta$.

We define the masked latent distance between a predicted state $\hat{s}$ and a target state s:

$$\ell(\hat{s}, s; m) = \frac{\sum_j m_j \, \|\mathcal{N}(\hat{s}_j) - \mathcal{N}(s_j)\|_2^2}{\sum_j m_j}, \tag{5}$$

where $\mathcal{N}$ applies layer normalization across feature channels, j indexes spatial patch tokens, and $m_j \in \{0, 1\}$ selects patches within the expanded organ region. For a patient with $K \geq 2$ observations at $t_0 < \cdots < t_{K-1}$, the paired prediction loss is

$$\mathcal{L}_{\text{pair}} = \frac{1}{K-1} \sum_{r=0}^{K-2} \ell(\hat{s}_{r+1|r}, s_{r+1}; m_r), \qquad \mathcal{L}_A = \mathcal{L}_{\text{pair}}, \tag{6}$$

where m_r is defined in the current organ CT crop's spatial reference and each prediction starts from the observed CT latent state s_r. Gradients update θ and ψ through the prediction branch. The target CT encoder updates only through EMA, while the text encoder remains frozen. The training objective is averaged over patients. Registration and mask construction are detailed in Appendix B.

3.2.2 Stage B: Dynamics Learning

In this stage, we freeze the adapted CT encoder as $E_\star$, stop EMA updates, and train only the latent predictor F_ψ using paired and chained training. The paired prediction loss follows Eq. (6), with both observed and target CT latent states constructed using $E_\star$.

For chained training, all follow-up CT scans are registered to the baseline spatial reference so that target and predicted CT latent states share corresponding spatial patch tokens. We define $s_r = E_\star(\tilde{x}_r)$ and $h_r = T_\omega(e_r)$, where $\tilde{x}_r$ denotes the registered organ CT crop at t_r, aligned to the baseline organ CT crop x_0 and $\tilde{x}_0 = x_0$.

Recursive rollout is defined by

$$\hat{s}_{r+1|0} = F_\psi(\hat{s}_{r|0}, h_r, \xi_r), \qquad \hat{s}_{0|0} = s_0. \tag{7}$$

Only s_0 initializes recursive rollout, and subsequent state transitions use predicted intermediate states. ξ_r specifies acquisition metadata and the transition interval for the corresponding state transition. Follow-up CT scans are used only to construct target CT latent states.

Using the same masked latent distance as in Stage A, the chained prediction loss compares each predicted CT latent state with its target CT latent state:

$$\mathcal{L}_{\text{chain}} = \frac{1}{K-1} \sum_{r=1}^{K-1} \ell(\hat{s}_{r|0}, s_r; m_0), \qquad \mathcal{L}_B = \mathcal{L}_{\text{pair}} + \lambda_{\text{chain}} \mathcal{L}_{\text{chain}}, \tag{8}$$

where m_0 is the expanded organ mask in the baseline spatial reference. Gradients propagate through the full prediction chain, allowing later errors to refine earlier state transitions and predicted intermediate states.

3.2.3 Stage C: Clinical Reader Adaptation

The visual bridge maps CT latent states to tokens for a Qwen3.5-9B clinical reader, which also receives a task prompt containing the task question and clinical context. The base LLM weights ϕ remain frozen, while QLoRA parameters α are trainable. Both clinical tasks share the clinical reader and use task-specific JSON answer formats.

We define the answer-text loss as

$$\ell_{\text{ans}}(s, q, u) = -\frac{1}{M} \sum_{j=1}^{M} \log p_{\phi,\alpha}\left(u_j \mid s, q, u_{<j}\right), \qquad M = |u|, \tag{9}$$

where s is the CT latent state, q is the task prompt, and $u = (u_1, \ldots, u_M)$ is the reference answer sequence. Here, $u_{<j}$ denotes the reference answer tokens preceding position j.

Observed CT latent state warm-up. We bypass the latent predictor and train only the clinical reader parameter α on observed CT latent states. For a patient with K observations, $s_r = E_\star(x_r)$ is the observed CT latent state at observation r.

We define

$$\mathcal{L}_{\text{warm}} = \frac{1}{K} \sum_{r=0}^{K-1} \ell_{\text{ans}}(s_r, q_r, u_r). \tag{10}$$

The CT encoder remains frozen, and the latent predictor and text encoder are not used.

Predicted-state reader adaptation. We then optimize α using predicted CT latent states obtained by single-step forecasting, $\hat{s}_{r+1|r} = F_\star(s_r, h_r, \xi_r)$, starting from observed CT latent states $s_r = E_\star(x_r)$, with clinical text tokens $h_r = T_\omega(e_r)$:

$$\mathcal{L}_C = \frac{1}{K-1} \sum_{r=0}^{K-2} \ell_{\text{ans}}\big(\hat{s}_{r+1|r}, q_{r+1|r}, u_{r+1}\big). \tag{11}$$

Here, $q_{r+1|r}$ asks about the outcome at t_{r+1} using only that available clinical context and numerical and categorical conditioning inputs. The reference answer u_{r+1} provides follow-up supervision and is excluded from all conditioning inputs. $E_\star$, T_ω, $F_\star$, and ϕ remain frozen.

Both objectives are averaged over patients, with task losses averaged within each CT scan or transition. Reference answers specify lesion diameter or binary liver metastasis status. At inference, recursive rollout propagates predicted CT latent states using Eq. (2). Stage C uses only reference answers for supervision, with no future-state supervision.

4 Experiments

4.1 Longitudinal Forecasting Benchmark

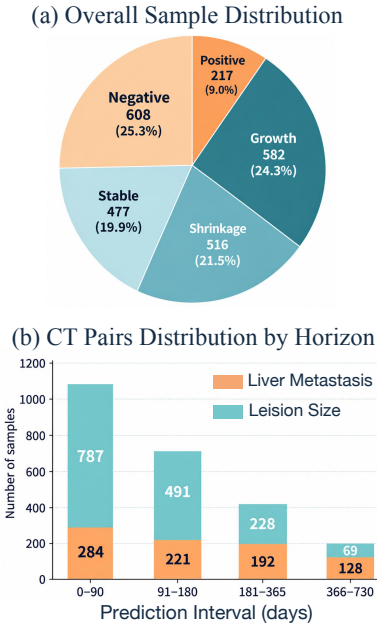


Figure 3: Dataset composition and category distribution of the two tumor forecasting tasks.

Dataset. We curate a longitudinal oncology dataset by linking longitudinal CT scans with radiology reports and EHRs. After restricting forecast horizons to at most 730 days, the dataset includes 1,344 patients with 3,606 CT scans, 2–10 CT scans per patient, and 11,728 structured clinical entries covering clinical history and interval treatment (see Fig. 1(b,c) for category distributions). We construct 2,400 forecasting examples across the liver and pancreas, including 1,575 for lesion size forecasting, and 825 for liver metastasis forecasting (Fig. 3(a,b)). Training and test sets are patient-disjoint, containing 2,158 and 242 forecasting examples, respectively.

Forecasting tasks and evaluation metrics. We introduce a longitudinal benchmark with two forecasting tasks (Fig. 1 (a)). The lesion task combines *lesion diameter forecasting* with three category classification (Shrinkage, Stable, Growth), while *liver metastasis forecasting* predicts binary liver metastasis status at follow-up. For liver metastasis forecasting, all examples have no documented liver metastasis at baseline, so the task specifically evaluates the emergence of new liver metastasis at follow-up. Growth requires a relative increase of at least 10% and an absolute increase of at least 2 mm. Shrinkage uses symmetric thresholds, and otherwise Stable. Each forecast uses the current organ CT crop, clinical history, interval treatment, forecast horizon, and acquisition metadata. For liver metastasis forecasting, we report AUPRC, sensitivity (Sens.), specificity (Spec.), and balanced accuracy (BAcc). For lesion size change classification, we report three-class BAcc over forecast horizons of 0–90, 91–180, 181–365, and 366–730 days, together with overall BAcc.

4.2 Experimental Setup

Baselines. We compare ***TumorCast*** with rendered-slice VLM baselines and text-only baselines. The VLM baselines include commercial models GPT-6 Astra (OpenAI, 2026b), GPT-5.6 Sol (OpenAI, 2026a), Claude Opus 5 (Anthropic, 2026), and Gemini 3.8 Flash (Google DeepMind, 2026); open-weight Qwen3.5-9B/27B (Team, 2026); and medical VLM MedGemma 1.5-4B (Sellergren et al., 2026). Open-weight VLMs are evaluated zero-shot or with task-specific fine-tuning. Text-only baselines include an adapted Clin-JEPA text-state model (Yang et al., 2026) and Qwen3.5-9B/27B. Rendered-slice VLM baselines receive reproducibly rendered slices from the current organ

Table 1: Comparison of ***TumorCast*** with rendered-slice VLM and text-only baselines on our tumor forecasting benchmark. AUPRC measures how well the model identifies future liver metastasis status across different decision thresholds. Lesion size change BAcc is reported by forecast horizon in days (n in parentheses). Mean BAcc = (liver metastasis BAcc + lesion size change BAcc)/2. ZS: zero-shot; SFT: supervised fine-tuning; Staged: Representation–Dynamics–Readout curriculum. * denotes re-implemented methods for fair comparison. **Bold** and underlined text indicates the best and second-best performance, respectively.

Model	**Setting**	**Liver Metastasis Forecasting**				**Lesion Size Change Classification** (BAcc ↑)					**Mean BAcc ↑**
		AUPRC ↑	Sens. ↑	Spec. ↑	BAcc ↑	0–90 d	91–180 d	181–365 d	366–730 d	BAcc ↑	
Commercial rendered-slice VLM baselines: CT slices + clinical context											
GPT-5.6 Sol	ZS	45.75	8.70	100.00	54.35	40.67	37.20	45.96	52.38	42.09	48.22
GPT-6 Astra	ZS	51.25	4.35	100.00	52.17	48.24	45.33	48.91	50.48	<u>48.42</u>	50.30
Claude Opus 5	ZS	52.05	13.04	96.49	54.77	44.42	35.35	26.94	45.71	40.12	47.45
Gemini 3.8 Flash	ZS	49.41	21.74	98.25	59.99	36.61	42.70	31.48	41.90	39.55	49.77
Open-weight rendered-slice VLM baselines: CT slices + clinical context											
MedGemma 1.5-4B	ZS	38.55	30.43	64.91	47.67	37.15	31.37	25.93	33.33	32.93	40.30
Qwen3.5-9B	ZS	45.99	21.74	96.49	59.12	32.79	36.13	42.42	40.00	36.04	47.58
Qwen3.5-27B	ZS	51.66	30.43	80.70	55.57	35.09	32.75	28.96	56.19	35.02	45.30
MedGemma 1.5-4B	SFT	38.72	65.22	54.39	59.80	30.36	39.94	26.94	56.19	34.88	47.34
Qwen3.5-9B	SFT	66.23	65.22	84.21	74.71	34.12	40.86	31.99	40.00	36.25	55.48
Qwen3.5-27B	SFT	70.65	95.65	54.39	75.02	34.91	34.60	37.37	46.67	35.75	55.39
Text-only baselines: clinical context only, no CT											
Clin-JEPA*	SFT	66.83	60.87	85.96	73.42	40.42	30.38	30.09	54.29	36.41	54.92
Qwen3.5-9B	SFT	64.72	95.65	52.63	74.14	30.55	38.39	30.56	44.76	33.75	53.95
Qwen3.5-27B	SFT	<u>71.17</u>	73.91	82.46	<u>78.18</u>	44.06	33.85	21.13	56.19	37.17	57.68
TumorCast variants: native 3D organ CT crops + clinical context											
TumorCast-9B	Staged	63.10	70.00	84.31	77.16	44.59	34.78	27.78	50.00	42.39	<u>59.78</u>
TumorCast-27B	Staged	**71.43**	75.00	82.35	**78.68**	52.24	46.01	39.90	67.62	**48.66**	**63.67**

CT crop, whereas ***TumorCast*** processes the native 3D organ CT crop. All methods receive the same task prompt including clinical history, interval treatment, forecast horizon, and task question.

Implementation details. We initialize the 3D CT encoder from M3D (Bai et al., 2024) and process $256 \times 256 \times 32$ CT views at $1 \times 1 \times 3$ mm spacing. A two-layer, width-256 AdaLN latent predictor produces predicted CT latent states, which a query-attention visual bridge maps to 32 visual tokens for Qwen3.5-9B or 128 for Qwen3.5-27B. Both readers are fine-tuned using 4-bit QLoRA with rank 16 and scaling parameter $\alpha = 32$. We use AdamW with learning rates of 10^{-5} for CT encoder adaptation and 10^{-4} for the latent predictor and clinical readout. See Appendix B for details.

4.3 Main Forecasting Results

Overall performance. Table 1 shows that ***TumorCast***-27B achieves the best aggregate result with 63.67% Mean BAcc, and the highest AUPRC and BAcc for liver metastasis forecasting, indicating superior positive-case ranking and more balanced discrimination between patients with and without future liver metastasis. By jointly training the CT encoder and latent predictor to align longitudinal CT latent states, both ***TumorCast*** variants substantially improve lesion size change classification over the Qwen3.5 rendered-slice VLM baselines, with the 27B model gaining about 13 BAcc points (48.66% vs. 35.75%).

Difficulty of lesion size change classification. Lesion size change classification remains the harder task, with a best overall BAcc of 48.66%. Rendered-slice VLM baselines MedGemma and Qwen3.5 stay near chance before and after SFT (32.93–36.25%), suggesting limited preservation of 3D geometry and cross-time lesion correspondence which are required to detect subtle longitudinal changes. A complementary contrast emerges from the text-only baselines: Qwen3.5-27B nearly matches ***TumorCast***-27B on liver metastasis BAcc (78.18% vs. 78.68%) but trails by 11.49 points on lesion size change BAcc. This suggests that clinical context is highly informative for liver metastasis forecasting, whereas lesion size change classification depends more strongly on spatial CT latent states learned from longitudinal CT scans.

Table 2: Ablations of ***TumorCast***. Single-step and Rollout denote single-step forecasting and recursive rollout, respectively. Mean BAcc = (liver metastasis BAcc + Single-step BAcc + Rollout BAcc)/3. S@10mm is the percentage of predictions with absolute diameter error at most 10 mm. **Bold** and underlined text indicates the best and second-best performance, respectively.

Model Variant	Liver Metastasis Forecasting			Lesion Diameter / Size Change				Mean BAcc ↑
	Sens. ↑	Spec. ↑	BAcc ↑	Single-step S@10mm ↑	Single-step BAcc ↑	Rollout S@10mm ↑	Rollout BAcc ↑	
TumorCast-27B (Full)	75.00	82.35	**78.68**	66.08	**48.66**	55.56	**50.41**	**59.25**
Bypass latent predictor	70.00	84.31	77.16	61.99	43.13	55.56	49.02	56.44
w/o clinical text conditioning	65.00	50.98	57.99	62.57	31.66	**57.78**	38.95	42.87
w/o chained prediction loss	75.00	76.47	75.74	**68.42**	42.27	**57.78**	43.37	53.79
Replace clinical reader with task head	35.00	76.47	55.74	67.84	31.46	53.33	35.71	40.97

Horizon-dependent behavior. For lesion size change classification, the BAcc of ***TumorCast***-27B progressively decreases across the 0–90, 91–180, and 181–365-day forecast horizon groups from 52.24% to 39.90%, before rising sharply to its highest value of 67.62% at 366–730 days. Our case analysis suggests that lesions in the longest-horizon group more often remain stable or show growth or shrinkage that clearly exceeds the classification thresholds, making these trajectories easier to distinguish. In contrast, subtle changes over shorter forecast horizons may be harder to separate from measurement variability.

4.4 Ablation Study

Table 2 evaluates the contributions of four major components of TumorCast: the latent predictor, clinical text conditioning, chained training, and the clinical reader.

Contribution of the latent predictor. Table 2 evaluates the contribution of each major component of ***TumorCast***. The full model achieves the highest BAcc for liver metastasis forecasting (78.68%) and lesion size change classification under both single-step forecasting and recursive rollout (48.66% and 50.41%). Bypassing the latent predictor reduces single-step lesion size change BAcc from 48.66% to 43.13% and liver metastasis BAcc from 78.68% to 77.16%, showing that explicitly predicting the future CT latent state improves downstream prediction beyond directly reading the observed CT latent state.

Importance of clinical text conditioning. Removing clinical text conditioning causes the largest degradation in liver metastasis forecasting: specificity falls from 82.35% to 50.98%, and BAcc decreases from 78.68% to 57.99%. Single-step lesion size change BAcc also drops to 31.66%, close to the 33.3% chance reference. These results indicate that clinical history and interval treatment provide essential clinical context for modeling tumor evolution from the observed CT latent state.

Effect of chained training and the clinical reader. Removing the chained prediction loss reduces lesion size change BAcc under single-step forecasting and recursive rollout from 48.66% and 50.41% to 42.27% and 43.37%, despite slightly improving S@10mm. This suggests that chained training improves lesion size change classification under recursive rollout as well as single-step forecasting. Replacing the clinical reader with a task head further lowers liver metastasis BAcc from 78.68% to 55.74% and single-step lesion size change BAcc from 48.66% to 31.46%, demonstrating the value of the clinical reader for integrating predicted CT latent states with clinical context.

5 Conclusion

TumorCast reframes tumor forecasting as CT latent state prediction in CT latent space. Follow-up CT scans provide target CT latent states for future-state supervision, while clinical context and transition interval condition each state transition. Recursive rollout propagates predicted CT latent states that support multiple predicted outcomes, avoiding both direct outcome prediction and future CT scan reconstruction. The learned latent dynamics encapsulate biologically plausible tumor progression patterns, decoupling image appearance from underlying disease evolution. This paradigm enables flexible downstream prediction tasks without retraining the core dynamics model. These results support latent dynamics as a promising direction for cancer world models of tumor evolution.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing, grammar correction, and improving the clarity of manuscript writing. The research ideas, methodological design, experimental result analysis were conducted by the authors.

ETHICS STATEMENT

This study used hospital-held clinical imaging and electronic health record data and was reviewed and approved by the relevant institutional Ethics Committee. Data collection and use complied with applicable laws, regulations, and ethical standards. All data were de-identified prior to use, with personally identifiable information removed. Access to the data was restricted to authorized personnel, and appropriate safeguards, including access controls and secure data handling, were implemented to protect patient privacy. The data were used solely for the purposes of this research.

REPRODUCIBILITY STATEMENT

The model architecture, training curriculum, preprocessing pipeline, hyperparameters, and evaluation protocols are described in the main paper and appendix. We will release the source code and training configurations upon publication to support reproducibility of ***TumorCast*** and the experimental analysis.

REFERENCES

Anthropic. Claude Opus 5 system card. `https://www.anthropic.com/claude-opus-5-system-card`, 2026. API model `claude-opus-5`, accessed Sep. 25, 2026.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*, volume 2026, pp. 89791–89807, 2026.

Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, 15(2):81–94, 2018.

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 preprint arXiv:2512.08029*, 2025.

Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

Google. Medgemma 1.5 model card. Official model documentation, 2026.

Google DeepMind. Gemini 3.8 Flash model card. `https://deepmind.google/models/model-cards/gemini-3-8-flash/`, 2026. API model `gemini-3.8-flash`, accessed Sep. 25, 2026.

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

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

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Zhifan Jiang, Dong Yang, Vishwesh Nath, Abhijeet Parida, Nishad P Kulkarni, Ziyue Xu, Daguang Xu, Syed Muhammad Anwar, Holger R Roth, and Marius George Linguraru. Lumen: Longitudinal multi-modal radiology model for prognosis and diagnosis. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2026.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, et al. Foresighta generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022. URL `https://openreview.net/forum?id=BZ5a1r-kVsf`.

Changsun Lee, Sangjoon Park, Cheong Il Shin, Woo Hee Choi, Hyun Jeong Park, Jeong Eun Lee, and Jong Chul Ye. Read like a radiologist: Efficient vision-language model for 3d medical imaging interpretation. *Medical Image Analysis*, pp. 104077, 2026.

Sijing Li, Zhongwei Qiu, Jiang Liu, Wenqiao Zhang, Tianwei Lin, Yihan Xie, Jianxiang An, Boxiang Yun, Chenglin Yang, Jun Xiao, et al. Tumorchain: Interleaved multimodal chain-of-thought reasoning for traceable clinical tumor analysis. In *International Conference on Learning Representations*, volume 2026, pp. 119790–119819, 2026.

Tianwei Lin, Zhongwei Qiu, Wenqiao Zhang, Jiang Liu, Yihan Xie, Mingjian Gao, Zhenxuan Fan, Zhaocheng Li, Sijing Li, Zhongle Xie, et al. Omnict: Towards a unified slice-volume lvlm for comprehensive ct analysis. *arXiv preprint arXiv:2602.16110*, 2026.

Massimiliano Mantegna, Elena Mulero Ayllón, Alice Natalina Caragliano, Francesco Di Feola, Claudia Tacconi, Michele Fiore, Edy Ippolito, Carlo Greco, Sara Ramella, Philippe C Cattin, et al. Longitudinal nsclc treatment progression via multimodal generative models. *arXiv preprint arXiv:2603.06147*, 2026.

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

James PB O'connor, Eric O Aboagye, Judith E Adams, Hugo JWL Aerts, Sally F Barrington, Ambros J Beer, Ronald Boellaard, Sarah E Bohndiek, Michael Brady, Gina Brown, et al. Imaging biomarker roadmap for cancer studies. *Nature reviews Clinical oncology*, 14(3):169–186, 2017.

OpenAI. GPT-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6, 2026a. GPT-5.6 Sol; API model gpt-5.6-sol, accessed Sep. 25, 2026.

OpenAI. GPT-6 Astra system card. https://deploymentsafety.openai.com/gpt-6-astra, 2026b. API model gpt-6-astra, accessed Sep. 25, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

Jiangyuan Wang, Xuyong Chen, Junwei He, Xu Xu, Shasha Xie, and Fuman Han. Chronomedical-world: A medical world model for learning patient trajectories from longitudinal care data. *arXiv preprint arXiv:2605.21963*, 2026.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 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.

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

Yixuan Yang, Mehak Arora, Ryan Zhang, Baraa Abed, Junseob Kim, Tilendra Choudhary, Md Hassanuzzaman, Kevin Zhu, Ayman Ali, Chengkun Yang, et al. Clin-jepa: A multi-phase co-training framework for joint-embedding predictive pretraining on ehr patient trajectories. *arXiv preprint arXiv:2605.10840*, 2026.

Zhenyu Yi, Qiang Hu, Zhenhao Li, Jiaxuan Zhao, Yusong Sun, and Lichi Zhang. Carve: Cross-slice anisotropic reallocation of visual evidence for efficient 3d medical volume understanding. *arXiv preprint arXiv:2608.04515*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 79115–79135, 2025.

A Dataset Construction

A.1 Clinical Records and Temporal Alignment

Each forecasting example links a current organ CT crop, baseline clinical history, supplied interval treatment, and a forecast horizon to a follow-up outcome. We retain intervals of at most 730 days and lesion diameter forecasting examples concerning the liver or pancreas. Clinical history describes information available at the forecast origin; interval treatment is supplied as a conditioning scenario. Follow-up reports and measurements define targets and are excluded from prediction inputs. Liver metastasis forecasting examples concern a liver documented as negative at baseline.

The cohort comprises 1,344 patients, 3,606 CT scans, and 2,400 forecasting examples: 1,575 for lesion diameter forecasting and 825 for liver metastasis forecasting. Clinical context contains 9,375 clinical history entries and 2,353 treatment entries, deduplicated by patient and entry text. These are structured clinical entries, not counts of independent EHR documents. A CT scan pair can support multiple forecasting examples. Figure 4 summarizes longitudinal coverage and class composition.

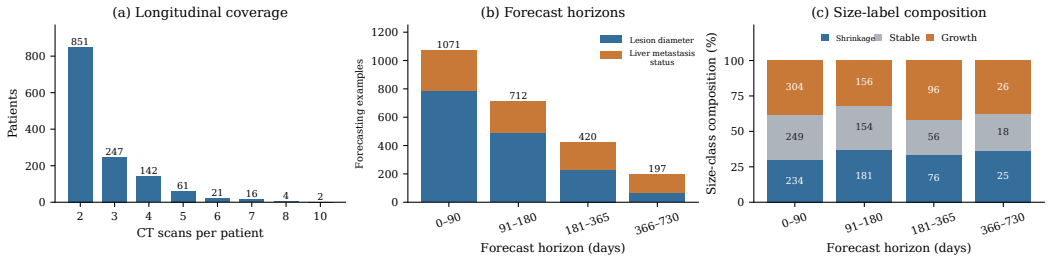


Figure 4: **Cohort composition across training and test data.** (a) Number of CT scans per patient. (b) Forecasting examples by forecast horizon and task. (c) Lesion size change class composition; segment labels give counts. These distributions describe the full curated cohort, not the test set alone.

B Implementation and Training Details

B.1 CT Preprocessing and Spatial Alignment

Organ segmentation. We obtain organ masks offline using TotalSegmentator v2.13.0 (Wasserthal et al., 2023).

Organ CT crops. We expand the selected organ bounding box by 10 mm and sample views at $1 \times 1 \times 3$ mm spacing. Each view contains $256 \times 256 \times 32$ voxels, spanning $256 \times 256 \times 96$ mm. Large organs use multiple views. CT values are linearly interpolated, clipped to $[-1000, 1000]$ HU, and mapped to $[0, 1]$; out-of-volume values use -1000 HU. Masks use nearest-neighbor interpolation. The organ CT crop retains surrounding anatomy inside the bounding box, rather than zeroing all non-organ voxels. The M3D encoder produces one global and 2,048 spatial patch tokens of width 768 per view.

Longitudinal registration. For paired training, the current CT scan defines the fixed reference and the follow-up CT scan is the moving image. The archived target pipeline uses organ-shape registration in SimpleITK. The saved transform maps a fixed physical point to its moving-image sampling location. The follow-up CT scan is resampled onto the current CT scan's view grids with linear interpolation, retaining the current organ CT crop bounds and padding.

B.2 Latent Predictor

The latent predictor projects CT latent state tokens from 768 to 256 dimensions and adds learned position embeddings. Two Transformer blocks each apply self-attention, gated text cross-attention,

Table 3: Stage-wise training schedule used in our experiments. The stage order, trainable modules, and supervision define the training curriculum. E: CT encoder; F: latent predictor; B: visual bridge; α: clinical reader LoRA. The target CT encoder is updated by EMA only.

Stage	Substage	Trainable	Input and supervision
Stage A Representation adaptation	Joint adaptation	E, F	Current organ CT crops are aligned with EMA-encoded follow-up CT latent states using masked latent distance.
	Alignment	F	The online CT encoder is frozen while the EMA target encoder approaches it, followed by hard synchronization.
Stage B Dynamics learning	Predictor finalization	F	Single-step paired training followed by two-step chained training.
Stage C Clinical reader adaptation	Text warm-up	α	Clinical text, baseline measurement, and forecast horizon are mapped to structured answers.
	Observed-state warm-up	B, α	Follow-up CT latent states are mapped to clinical answers with the latent predictor bypassed.
	Predicted-state adaptation	B, α	Predicted CT latent states are mapped to clinical answers with the CT encoder and predictor frozen.

and a feed-forward network. Each attention module uses four heads; the feed-forward expansion is four. Frozen Qwen3.5-2B clinical text tokens condition cross-attention through a 2048 to 256 projection.

Numerical and categorical conditioning includes baseline diameter, transition interval, organ identity, and acquisition phase. Diameter and transition interval use both scaled linear and logarithmic features. A two-layer MLP and a four-category acquisition-phase embedding produce the AdaLN condition. Acquisition phase is randomly set to unknown with probability 0.1 during training. AdaLN predicts scale and shift for the self-attention and feed-forward branches. Its final projection is zero-initialized. The output is a residual update. A zero transition interval returns the input CT latent state.

B.3 Clinical Reader and Visual Bridge

A learned query-attention visual bridge maps predicted CT latent states into the Qwen3.5 language embedding space. It projects each 2049×768 view to width 256, adds position embeddings, and uses four-head attention to pool the sequence into Q learned queries. Corresponding query outputs are averaged across views, normalized, and projected to the clinical reader dimension. The implementation scales the resulting visual prefix by 0.02 before concatenating it with text embeddings. We use $Q = 32$ for the 9B clinical reader and $Q = 128$ for the 27B clinical reader.

B.4 Stage-wise Optimization

Table 3 summarizes the training schedule used in our final experiments. For clarity, the main method section focuses on the optimization steps essential to the Representation–Dynamics–Readout curriculum and omits auxiliary implementation details, such as the reader/interface warm-up procedure. We provide the complete stage-wise schedule here. Stage C updates the QLoRA parameters α and visual bridge interface of the clinical reader, while the CT encoder and latent predictor remain frozen.

C Baselines and Task Prompts

C.1 CT Rendering Protocol

Rendered-slice VLM baselines receive deterministic slices from the current CT scan. The protocol uses eight organ CT crop slices for lesion diameter forecasting and four full-field plus eight liver CT crop slices for liver metastasis forecasting. Slice order is fixed within each request. Clinical context, interval treatment, and forecast horizon accompany the images.

C.2 Text-Only Baselines

Text-only Qwen models receive clinical history, supplied interval treatment, organ, baseline measurement, and forecast horizon, without CT pixels or visual-prefix tokens. They are evaluated using the same task endpoints.

The adapted Clin-JEPA baseline encodes clinical history, current tumor state, interval treatment, and transition interval using a shared Qwen3.5-9B text encoder with rank-16 QLoRA. A six-layer, width-512 block-causal Transformer predicts the future text-state embedding. Targets use measured future tumor states during training. Training comprises one text-initialization epoch, followed by predictor warm-up (2 epochs), encoder–predictor co-training (2), alignment (1), hard synchronization, and finalization (3). Only co-training updates the online encoder adapters; the target adapters follow EMA with momentum 0.996. Task-specific MLP probes read the frozen representations.

D Evaluation Metrics and Task Prompts

D.1 Diameter-Derived Change Labels

For a nonnegative reference diameter d and future diameter y, let $\tau(d) = \max(0.1d, 2\,\mathrm{mm})$. We define

$$c(d, y) = \begin{cases} \text{Growth}, & y - d \geq \tau(d), \\ \text{Shrinkage}, & d - y \geq \tau(d), \\ \text{Stable}, & \text{otherwise}. \end{cases} \tag{12}$$

Both boundaries are inclusive.

D.2 Single-Step Forecasting and Recursive Rollout Protocols

Single-step forecasting predicts t_{r+1} from the observed CT latent state at t_r. Its reference class is $c(d_r, d_{r+1})$. Recursive rollout initializes once at t_0 and recursively predicts two CT latent states. The latent predictor receives each transition interval; the clinical reader receives the forecast horizon from t_0. In the archived evaluation, the final reference class is $c(d_0, d_2)$ and the diameter-derived prediction is $c(d_0, \hat{d}_2)$. Both are relative to the original baseline.

D.3 Metric Definitions

For n lesion diameter predictions, S@10mm is the percentage with absolute diameter error at most 10 mm:

$$\mathrm{S@10mm} = \frac{100}{n} \sum_i \mathbf{1}\{|\hat{y}_i - y_i| \leq 10\,\mathrm{mm}\}. \tag{13}$$

For K supported classes, balanced accuracy is mean recall:

$$\mathrm{BAcc} = \frac{100}{K} \sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}. \tag{14}$$

Here $K = 3$ for lesion size change classification and $K = 2$ for liver metastasis forecasting. Binary sensitivity (Sens.) and specificity (Spec.) are positive- and negative-class recall at probability threshold 0.5. All reported lesion size change strata contain all three classes.

The archived task prompts below retain their original terminology and output schemas for reproducibility.

D.4 INFERENCE PROMPTS

The following templates retain the task instructions and output schema of the archived inference code. Bracketed fields denote sample-specific inputs; Q is 32 or 128. Actual inputs serialize the baseline clinical fields and visual provenance as JSON. Prompt prefixes are shortened here only to avoid repeating the same provenance text.

Table 4: Shared context and task-specific inference instructions. Target answers are never included in the input fields.

System instruction Respect the CT source and time stated in visual_context and the task question. Baseline clinical measurements refer to day 0. Model-predicted future features are not observed scans. Return only the requested JSON.
Structured context History: [baseline history]. Treatment: [supplied interval treatment]. Organ: [organ]. Current diameter: [baseline diameter in mm, for size]. Horizon: [days]. Current phase: [phase]. Visual source: [predicted future CT features / no CT input].

Lesion-size question
The Q visual prefix tokens are MODEL-PREDICTED future CT features for day [horizon] after the original baseline CT (day 0). They are not features from an observed future CT. Use the baseline clinical information and supplied interval treatment scenario to predict the outcome at this horizon. Predict the future diameter in mm of the same target lesion. Classify size change relative to current_diameter_mm at the original baseline day 0. Shrink: relative decrease $\geq 10\%$ AND absolute decrease ≥ 2 mm; growth: relative increase $\geq 10\%$ AND absolute increase ≥ 2 mm; otherwise stable. Return only JSON:
`{"future_diameter_mm": <number>,`
`"size_change": "shrink|stable|growth"}`

Liver-metastasis question Use the baseline clinical information and supplied interval treatment scenario to predict the outcome at day [horizon]. The liver at baseline day 0 is known negative for tumor presence. Predict liver tumor presence at the specified future day. Return only JSON: `{"future_liver_metastasis": <0 or 1>}`
Full-rollout addition This is the second predicted future endpoint; horizon_days is the TOTAL elapsed time from the original baseline. Only the original baseline CT was observed. The intermediate state and previous diameter are model predictions; no observed intermediate CT is supplied. Later-interval treatment is unknown; no treatment decisions from the intermediate visit are supplied.

E SUCCESSFUL AND FAILED FORECASTS

Table 5 presents one correctly classified example from each lesion size change class, selected by median absolute error among correct and internally consistent predictions. S1 predicts shrinkage from 17 to 10 mm, compared with an observed endpoint of 14 mm. S2 predicts a stable lesion of 13 mm, compared with 12 mm observed. S3 predicts pancreatic growth from 22 to 31 mm, compared with 25 mm observed. These examples illustrate correct change-direction prediction despite residual diameter error.

Table 5: De-identified numerical examples from the test predictions. Diameters are in mm. S1–S3 are the median-error examples among correct predictions in each class (requiring agreement between direct and diameter-derived classes). F1 and F2 are the largest- and smallest-error direct-class failures. These examples are descriptive, not prevalence estimates.

Case	Organ	Forecast horizon (days)	Baseline	True	Pred.	True class	Pred. class
S1	Liver	149	17	14	10	Shrinkage	Shrinkage
S2	Liver	422	13	12	13	Stable	Stable
S3	Pancreas	56	22	25	31	Growth	Growth
F1	Liver	592	20	90	15	Growth	Shrinkage
F2	Pancreas	64	39	35	36	Shrinkage	Stable