From Exams to Trajectories: Neural Controlled Differential Equations for Longitudinal Mammography Vision-Language Pretraining

Zijun Sun[1 [0009−0004−1848−6966]], Enze Ge[2,3 [0009−0009−6787−4320]] Solveig Thrun[1 [0009−0006−7349−9449]], and Michael Kampffmeyer[1,4 [0000−0002−7699−0405]]

[1] Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø, Norway zijun.sun@uit.no
[2] Norwegian Institute of Bioeconomy Research (NIBIO), Ås, Norway
[3] Norwegian University of Life Sciences (NMBU), Ås, Norway
[4] Norwegian Computing Center, Oslo, Norway

Abstract. Longitudinal analysis of a patient's screening history is fundamental to mammography interpretation. Yet existing deep learning models struggle with the irregular, continuous-time nature of screening data, where patient histories involve unevenly timed multi-view exams paired with evolving textual reports. To address this gap, we introduce **Dynamo**, a novel multi-modal pretraining **dyna**mic framework for longitudinal **m**amm**o**graphy with two key innovations to model exam timelines as latent trajectories, i.e., a coarse-fine grained exam-level temporal encoder based on Neural Controlled Differential Equations (Neural CDEs) and a Temporal Visual Question Answering (TVQA) loss for query-driven masked token prediction conditioned on temporal context. We perform comprehensive evaluations on downstream tasks including risk prediction using large-scale datasets (EMBED and CSAW-CC), BI-RADS assessment, and breast density classification. Across all benchmarks, Dynamo achieves overall gains over state-of-the-art vision-only and CLIP-style models, improving both calibration and temporal reasoning. On the EMBED dataset, representative gains include 11.81% relative reduction in Risk Prediction's Brier Score, 3.59% relative improvement in BI-RADS κ, and 2.39% relative improvement in BI-RADS AUC in zero-shot settings. Our code is available at this URL.

Keywords: Longitudinal Mammography · Dynamic Modeling · Vision-Language Pretraining.

1 Introduction

Breast cancer remains a leading cause of cancer-related mortality among women worldwide [1], motivating screening programs that depend on early, reliable detection. Full-field digital mammography (FFDM) is the primary screening modality due to its broad availability and demonstrated effectiveness [30]. Each screening exam typically includes two views per breast (CC and MLO), producing four images. In clinical practice, interpretation is inherently longitudinal:

radiologists compare current findings against prior mammograms and associated reports to identify subtle temporal changes, improving detection while reducing avoidable recalls.

Despite this workflow, most deep learning systems for mammography interpret a single multi-view exam in isolation [32,29,13]. Recent work has started to incorporate screening histories for risk prediction, with models such as MMT [25], LoMaR [15], and VMRA-MaR [27] using Transformer [28] or Mamba [10]-style backbones to process sequences of prior exams. While promising, current longitudinal designs leave two gaps. (1) Time is typically injected through discrete positional/temporal encodings over visits, which does not provide a continuous-time latent evolution and can be fragile under irregular inter-exam intervals, missing visits, and long gaps [4]. (2) Most approaches rely on imaging alone, leaving the accompanying radiology reports unused despite their central role in clinical comparison.

Separately, vision–language modeling has been explored for mammography by aligning images with free-text reports, often via CLIP-like objectives [23], improving representation learning and downstream performance [7,8]. These methods, however, operate at a single time point and therefore do not explicitly ground language in longitudinal visual evolution.

We propose *Dynamo*, a multimodal pretraining framework for longitudinal mammography that addresses both limitations. Dynamo models each patient's screening history as a continuous-time latent trajectory via a coarse-to-fine, exam-level Neural Controlled Differential Equation (Neural CDE) encoder [16], enabling principled handling of irregularly sampled timelines. To couple language with temporal visual context, we introduce a Temporal Visual Question Answering (TVQA) objective that conditions masked-token prediction on the learned trajectory, encouraging alignment between report language and image changes across exams. Our contributions are:

- **Continuous-time longitudinal encoder.** An exam-level Neural CDE temporal module that directly models irregular inter-exam intervals and yields robust patient trajectories without discretization or heuristic positional encodings.
- **Temporal VQA supervision.** A query-driven TVQA loss that aligns free-text descriptions with temporally evolving image features, promoting explicit reasoning over change across visits.
- **Unified multimodal pretraining for screening.** A scalable pretraining recipe that leverages multi-view mammograms and paired reports over patient histories, producing temporally grounded representations for downstream detection, risk prediction, and report understanding.

2 Method

Our proposed Dynamo shown in Fig. 1 explicitly models the dynamically obtained screening information by leveraging a Neural CDE (NCDE) [16] module to fuse the image information (four-view mammograms) and the radiology text

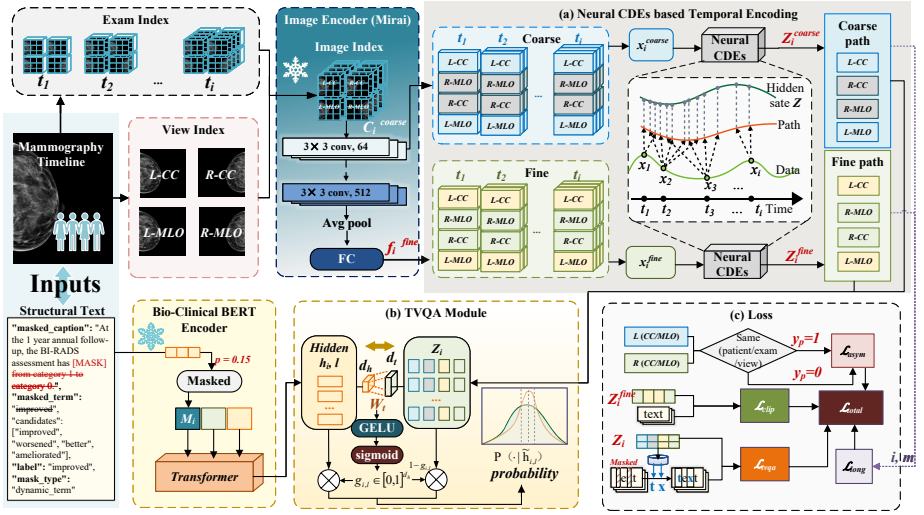


Fig. 1: Overview of **Dynamo**. (a) Coarse/fine Neural CDE trajectory encoder over irregular exam times; (b) TVQA trajectory-conditioned masked-token prediction with Bio-Clinical BERT; (c) auxiliary contrastive losses for multimodal regularization.

into a unified longitudinal representation. Dynamo applies a frozen Mirai encoder to each view, forming coarse exam level and fine view level streams. These streams define continuous control paths over exam times and drive parallel Neural CDEs, producing temporal embeddings that summarize patient history up to each visit. The resulting context conditions TVQA prediction of masked report terms, while asymmetry, longitudinal consistency, and image and text contrastive losses provide clinical and multimodal regularization..

Notation. $[a; b]$ denotes concatenation, $\odot$ denotes element-wise product, and $\mathrm{norm}(a) = a/\|a\|_2$.

2.1 Temporal Encoding via Neural CDEs

While exams are discrete, the underlying disease evolution is continuous and sampled at irregular intervals, we therefore model the patient's history as a continuous latent trajectory using a NCDE [16].

This contrasts with discrete-step models (e.g., RNNs, Transformers) that rely on heuristic positional encodings. The framework inherently uses real-time gaps as integration periods for the latent state, enabling it to learn nuanced temporal progression patterns from sparse and irregularly-sampled mammography exams.

Multi-view feature extraction. Each exam comprises four views: L-CC, R-CC, L-MLO, R-MLO (L/R = left/right breast). Following common mammography pipelines [31,15,27], we apply a shared, frozen Mirai image encoder [32] $E(\cdot)$ to each view. To capture multi-scale information, we extract features from two

different depths of the network. For each view $j \in \{1, \ldots, M\}$ ($M = 4$ views) of exam i, we obtain an early-layer feature $\mathbf{v}_{c,i}^{(j)}$ and a deeper-layer feature $\mathbf{v}_{f,i}^{(j)}$. We then derive two levels of representation per exam: (i) a *coarse* exam summary $\mathbf{c}_i \in \mathbb{R}^{d_c}$ formed by pooling the earlier-layer features across views, and (ii) a *fine* per-view representation $\mathbf{f}_i \in \mathbb{R}^{M \times d_f}$ formed by concatenating the deeper-layer features. Concretely,

$$\mathbf{c}_i \;=\; \tfrac{1}{M} \sum_{j=1}^{M} \mathbf{v}_{c,i}^{(j)}, \quad \mathbf{f}_i \;=\; \big[\, \mathbf{v}_{f,i}^{(1)}; \mathbf{v}_{f,i}^{(2)}; \ldots; \mathbf{v}_{f,i}^{(M)} \,\big]. \tag{1}$$

To align dimensions and control capacity we project the coarse and fine features with learned linear maps $\hat{c}_i = W_c c_i + b_c$ and $\hat{f}_i = W_f f_i + b_f$. We append timestamps as an extra channel and build continuous control paths $X_c(t)$ and $X_f(t)$ by linear interpolation over knots $\{(t_i, [\hat{c}_i; t_i])\}_{i=1}^{N}$ and $\{(t_i, [\hat{f}_i; t_i])\}_{i=1}^{N}$.

Neural CDE dynamics (per stream). For $s \in \{c, f\}$, a Neural CDE evolves a hidden state $\mathbf{z}_s(t) \in \mathbb{R}^{h_s}$ driven by the corresponding path $\mathbf{X}_s(t)$:

$$\mathbf{z}_s(t) \;=\; \mathbf{z}_s(t_1) + \int_{t_1}^{t} f_\theta^{(s)}\big(\mathbf{z}_s(\tau)\big)\, \mathrm{d}\mathbf{X}_s(\tau), \tag{2}$$

Where $\mathbf{z}_s(t_1) = \zeta_\theta^{(s)}\big(\mathbf{X}_s(t_1)\big)$ initializes the hidden state based on the first observation, and time is included as an explicit channel in the input path. We numerically integrate Eq. (2) forward with an adaptive Runge–Kutta solver (`dopri5`) [2,3] to obtain $\{\mathbf{z}_s(t_i)\}_{i=1}^{N}$.

Fusion at exam times. At each exam, we compute a time conditioned temporal embedding by fusing the per-stream states: $\mathbf{z}_i = [\mathbf{z}_c(t_i); \mathbf{z}_f(t_i)] \in \mathbb{R}^h$,

Contextual temporal embedding. We refer to $\mathbf{z}_i$ as the temporal embedding for exam i. By integrating up to t_i, $\mathbf{z}_i$ summarizes the entire imaging history through exam i and serves as a compact, time-conditioned context for downstream reasoning. In effect, $\mathbf{z}_i$ functions as an informed prior over likely findings at visit i given the observed trajectory.

2.2 Temporal Visual Question Answering (TVQA)

To facilitate longitudinal vision-language pretraining, we pair each exam i with a report token sequence $(w_{i,1}, \ldots, w_{i,L_i})$. Following the masked language modeling paradigm, we independently mask each token with probability p=0.15, yielding the masked index set $\mathcal{M}_i \subseteq \{1, \ldots, L_i\}$. Text is encoded with Bio-Clinical BERT [17], producing contextual states $\mathbf{h}_{i,l} \in \mathbb{R}^{d_h}$. Let $\mathbf{z}_i \in \mathbb{R}^{d_z}$ be the temporal embedding from Sec. 2.1.

Masked-token prediction with temporal context. We first align dimensions by projecting $\mathbf{z}_i$ to the text hidden size:

$$\tilde{\mathbf{z}}_i \;=\; W_t \mathbf{z}_i + \mathbf{b}_t, \quad W_t \in \mathbb{R}^{d_h \times d_z},\ \mathbf{b}_t \in \mathbb{R}^{d_h}. \tag{3}$$

For each masked position $l \in \mathcal{M}_i$, we form a joint input $\mathbf{u}_{i,l} = [\mathbf{h}_{i,l}, \tilde{\mathbf{z}}_i] \in \mathbb{R}^{2d_h}$. To selectively route information between textual grammar and visually

grounded findings, we capture complex nonlinear multimodal dependencies using an adaptive gate:

$$\mathbf{g}_{i,l} = \sigma\Big(W_{g2}\,\mathrm{GELU}\big(W_{g1}\,\mathbf{u}_{i,l} + \mathbf{b}_{g1}\big) + \mathbf{b}_{g2}\Big), \tag{4}$$

with $\mathbf{g}_{i,l} \in [0,1]^{d_h}$, $W_{g1} \in \mathbb{R}^{d_h \times 2d_h}$, $W_{g2} \in \mathbb{R}^{d_h \times d_h}$, and biases $\mathbf{b}_{g1}, \mathbf{b}_{g2} \in \mathbb{R}^{d_h}$. The gated fusion interpolates between modalities:

$$\tilde{\mathbf{h}}_{i,l} \;=\; \mathbf{g}_{i,l} \odot \mathbf{h}_{i,l} \;+\; \big(\mathbf{1} - \mathbf{g}_{i,l}\big) \odot \tilde{\mathbf{z}}_i, \tag{5}$$

where $\tilde{\mathbf{z}}_i$ is broadcast over l. When $\mathbf{g}_{i,l} \approx \mathbf{1}$ the model leans on text; $\mathbf{g}_{i,l} \approx \mathbf{0}$ it relies on the temporal visual context.

Prediction head and loss. A standard two-layer MLP prediction head (with LayerNorm and GELU) computes the vocabulary distribution $P(\cdot \mid \tilde{\mathbf{h}}_{i,l})$ from the final hidden state. The model is trained using the cross-entropy loss over the N masked tokens in the batch:

$$\mathcal{L}_{\mathrm{TVQA}} = -\frac{1}{N} \sum_{i} \sum_{l \in \mathcal{M}_i} \log P(w_{i,l} \mid \tilde{\mathbf{h}}_{i,l}). \tag{6}$$

This objective formulates each mask as a visual query for exam i, conditioned on its longitudinal history, to drive the reconstruction of visually grounded terms.

2.3 Asymmetry-aware contrastive loss

Bilateral symmetry provides a strong inductive bias in screening mammography, where unilateral deviations are clinically salient [18,5]. We encourage contralateral embeddings from the same exam and view to be close while separating mismatched pairs.

For exam i and view $v \in \{\mathrm{CC}, \mathrm{MLO}\}$, let $\mathbf{f}_{i,\mathrm{L},v}$ and $\mathbf{f}_{i,\mathrm{R},v}$ be ℓ_2-normalized image embeddings (after a projection head). Construct pairs $(\mathbf{a}_p, \mathbf{b}_p)$ with labels $y_p \in \{0,1\}$, where y_p=1 iff $(\mathbf{a}_p, \mathbf{b}_p) = (\mathbf{f}_{i,\mathrm{L},v}, \mathbf{f}_{i,\mathrm{R},v})$ for some (i,v), and y_p=0 otherwise. Negatives are sampled across patients and matched by view. With margin $\delta \in (0,1)$ and $c_p = \cos(\mathbf{a}_p, \mathbf{b}_p)$, the loss is

$$\mathcal{L}_{\mathrm{asym}} = \tfrac{1}{|\mathcal{A}|} \textstyle\sum_{p \in \mathcal{A}} \left[y_p(1 - c_p) + (1 - y_p)[c_p - \delta]_+ \right] \tag{7}$$

2.4 Auxiliary contrastive modules

We add two auxiliary terms to regularize the representation. (i) **Longitudinal contrastive consistency**($\mathcal{L}_{\mathrm{long}}$) **:** we fuse each view embedding with its temporal context, $r_{i,m} = \mathrm{norm}([f_{i,m}; z_i])$, and apply a multi-positive InfoNCE objective that treats all other images from the same patient as positives [21]. (ii) **Image–text alignment** ($\mathcal{L}_{\mathrm{clip}}$)**:** we align image and report embeddings with the symmetric CLIP contrastive loss [23].

Overall objective. The total pretraining loss is a weighted sum of the TVQA loss and the auxiliary losses:

$$\mathcal{L}_{\mathrm{total}} \;=\; \lambda_{\mathrm{tvqa}}\,\mathcal{L}_{\mathrm{TVQA}} + \lambda_{\mathrm{asym}}\,\mathcal{L}_{\mathrm{asym}} + \lambda_{\mathrm{long}}\,\mathcal{L}_{\mathrm{long}} + \lambda_{\mathrm{clip}}\,\mathcal{L}_{\mathrm{clip}}, \tag{8}$$

where each ($\lambda_\bullet \geq 0$) is a trade-off hyperparameter selected on a validation set.

3 Experiments

3.1 Dataset

We use two longitudinal cohorts: EMBED [14] and CSAW-CC [26]. All baselines are trained and/or evaluated on the same EMBED/CSAW-CC patient wise 70/10/20 train/val/test splits as Dynamo. Both contain multi-view FFDM screening timelines, with multiple exams available for ≈64% (EMBED) and ≈75% (CSAW-CC) of patients. EMBED provides pathology-verified outcomes; we generate de-identified structured pseudo-reports via deterministic slot filling from tabular annotations following [7], excluding post-hoc outcome fields to prevent leakage. CSAW-CC lacks textual records, so evaluation there uses vision-only inference.

3.2 Implementation details

We train **Dynamo** on 8× NVIDIA H200 GPUs for 30 epochs (batch size 256) with learning rate 2×10^{-3} (2-epoch warmup) decayed to 10^{-6}, weight decay 0.01, gradient-norm clipping at 1.0, and 8 data-loading workers. We optimize end-to-end with the multi-task loss in Eq. (8), using masked-token TVQA as the primary term and asymmetry, longitudinal, and image–text (CLIP) contrastive losses as auxiliary alignment; λ_{tvqa}=1.0 and λ_{asym}=λ_{long}=λ_{clip}=0.5.

Evaluation metrics and reporting. Following prior mammography risk modeling [32,5,27] and survival/classification evaluation [11,9,12,20], we report calibration (WMAE; sample-weighted) and probabilistic accuracy (Brier Score), ranking concordance (Harrell's C-index), discrimination (AUC at 1–5 y), and agreement (Cohen's κ). For table consistency, we express AUC, C-index, and κ in percentage points (%), and display WMAE and BS on the same scale.

Computational requirements. The final Dynamo model contains 121M parameters. Pretraining was performed using PyTorch FP16 distributed data parallel training on 8 NVIDIA H200 GPUs, requiring 132 total GPU and CPU hours. Inference on 115,810 test samples took approximately 0.3 h, corresponding to 0.0026 h per 1,000 samples.

3.3 Results on downstream tasks

Risk prediction and calibration (EMBED). Table 1 indicates Dynamo shows strong overall performance. It achieves the best calibration (lowest WMAE and BS) and the highest C-index (82.79%). For discrimination, Dynamo shows high long-horizon (3–5 year) AUCs, while VMRA-MaR [27] performs best at early (1–2 year) horizons. Overall, Dynamo shows better calibration, concordance, and long-horizon discrimination than the baselines.

BI-RADS classification. In Table 1, CLIP-style pretraining outperforms Vision-only in zero-shot. Dynamo shows the highest performance across zero-shot, linear, and full fine-tuning(FT) settings, outperforming MaMA [7]. Table 2 further

Table 1: **Risk and BI-RADS performance on EMBED.** We report calibration (WMAE ↓, BS ↓), ranking (C-index ↑), 1–5y risk AUCs, and BI-RADS AUC and κ under zero-shot / linear probing / full fine-tuning. Methods are grouped as *Vision Only*, *CLIP style*, and **Dynamo**. Bold/underline mark best/second-best.

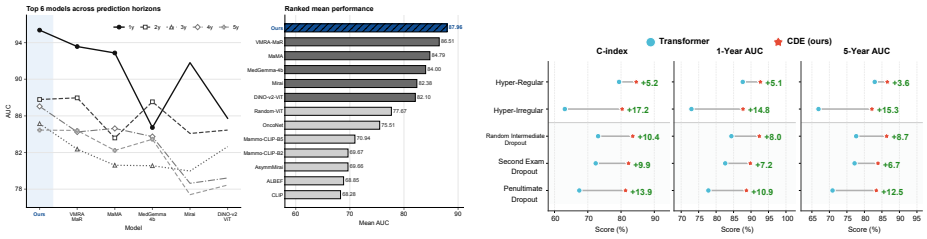
Model	WMAE	BS (%)	C-index	Risk Prediction (AUC)					BI-RADS AUC (%)			BI-RADS κ (%)		
				1y	2y	3y	4y	5y	Zero-shot	Linear	Full FT	Zero-shot	Linear	Full FT
Vision Only														
OncoNet [31]	2.35	34.97	71.71	81.36	78.99	74.10	73.26	66.63	–	59.83	61.25	–	57.65	58.13
Random-ViT [6]	2.23	32.62	72.82	82.29	78.00	75.19	74.18	72.24	–	60.45	61.78	–	58.20	60.52
AsymmMirai [5]	1.96	27.43	80.72	88.16	80.14	78.87	76.93	74.93	–	61.22	62.34	–	60.71	61.10
DiNOv2-ViT [22]	2.17	27.97	79.69	85.09	83.70	81.96	77.98	78.14	–	69.58	67.92	–	68.32	67.47
Mirai [32]	2.10	31.25	78.96	90.23	80.82	78.01	78.91	75.91	–	71.24	72.54	–	68.89	72.15
VMRA-MaR [27]	1.45	25.26	<u>82.14</u>	**93.65**	**87.30**	**84.28**	81.15	<u>84.38</u>	–	74.32	75.83	–	71.58	75.36
CLIP style														
CLIP [23]	1.92	26.46	74.26	80.90	75.37	59.61	61.54	61.00	58.12	73.71	72.15	55.61	72.40	71.72
ALBEF [19]	1.88	23.26	75.34	83.53	79.97	77.25	62.39	60.08	67.00	70.98	71.24	65.82	70.27	70.88
MedGemma-4b [24]	1.78	24.05	73.20	84.86	76.24	67.82	63.95	66.65	68.32	71.52	73.09	67.68	71.21	72.35
Mammo-CLIP-B2 [8]	2.09	37.50	72.32	71.73	64.08	69.36	67.40	62.10	55.31	66.58	69.20	52.72	66.03	69.00
Mammo-CLIP-B5 [8]	1.85	33.65	69.52	74.30	68.30	64.03	65.84	68.39	56.83	69.14	71.34	55.40	68.03	70.66
MaMA [7]	<u>1.42</u>	<u>15.32</u>	77.62	86.72	86.37	82.54	<u>81.84</u>	82.90	<u>75.24</u>	<u>78.33</u>	<u>78.12</u>	<u>73.26</u>	<u>77.48</u>	<u>77.57</u>
Ours														
Dynamo	**1.38**	**13.51**	**82.79**	<u>91.35</u>	<u>84.61</u>	**84.28**	**81.91**	**85.19**	**77.04**	**78.54**	**79.01**	**75.89**	**77.84**	**78.34**

Table 2: **Risk prediction** AUC (1y/5y) by BI-RADS density and BI-RADS classification performance (Full FT) on EMBED.

Model	C-index (%)				Risk Pred. AUC (1y) (%)				Risk Pred. AUC (5y) (%)				BI-RADS AUC and κ (%)							
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
Vision Only																				
OncoNet [31]	73.54	74.59	70.79	67.92	85.48	82.23	81.39	76.34	68.79	67.54	66.66	63.53	63.27	62.74	60.03	58.96	60.64	59.07	57.19	55.62
Random-ViT [6]	76.85	74.20	71.59	68.64	85.37	82.30	81.26	80.23	75.34	73.25	71.99	68.38	63.94	62.57	61.32	59.29	62.41	62.54	58.29	58.84
AsymmMirai [5]	80.56	78.93	82.79	80.60	<u>93.10</u>	89.82	86.34	83.38	76.79	74.74	74.38	73.81	64.68	63.92	61.71	59.05	64.10	63.81	61.26	55.23
DiNO-v2-ViT [22]	80.56	**81.52**	78.75	77.93	88.02	89.33	83.64	79.37	82.09	81.48	76.65	72.34	69.10	68.27	67.93	66.38	69.05	67.92	67.80	65.11
Mirai [32]	**82.77**	79.59	78.33	75.15	92.92	91.43	89.37	87.20	78.57	77.10	74.63	73.34	75.69	73.26	73.18	68.03	74.93	73.17	72.85	67.65
VMRA-MaR	<u>81.06</u>	81.09	<u>82.93</u>	<u>83.48</u>	**94.66**	<u>91.59</u>	**94.73**	**93.62**	83.42	<u>85.43</u>	<u>84.48</u>	<u>84.19</u>	76.28	77.31	74.51	75.22	75.98	77.19	73.62	74.65
CLIP style																				
CLIP [23]	77.96	75.46	73.26	70.36	84.83	83.17	79.88	75.72	64.96	60.09	61.07	57.88	71.29	72.25	72.09	72.97	70.94	72.14	71.85	71.95
ALBEF [19]	78.28	76.17	75.49	71.42	87.58	84.53	83.50	78.51	62.47	60.61	59.48	57.76	73.23	72.60	70.58	68.55	72.98	72.41	70.11	68.02
MedGemma-4b [24]	76.49	74.23	73.12	68.96	86.70	87.53	81.42	83.79	66.57	72.70	64.62	62.71	75.91	74.26	71.08	71.11	74.82	74.10	70.73	69.75
Mammo-CLIP-B2 [8]	75.08	74.27	71.47	68.46	74.56	75.83	68.02	68.51	63.85	63.14	62.40	59.01	70.19	70.07	68.37	68.17	69.81	69.97	68.32	67.90
Mammo-CLIP-B5 [8]	75.57	69.50	68.51	64.50	76.24	77.48	72.13	71.35	70.36	68.41	69.47	65.32	72.41	73.54	69.20	70.21	72.15	73.01	68.97	68.51
MaMA [7]	79.52	78.62	76.80	75.54	87.61	89.74	87.70	81.83	<u>83.74</u>	82.25	84.01	81.60	<u>79.01</u>	<u>79.24</u>	<u>78.16</u>	<u>76.07</u>	<u>78.11</u>	<u>78.24</u>	<u>77.80</u>	<u>75.77</u>
Ours																				
Dynamo	80.77	<u>81.50</u>	**84.05**	**84.84**	89.32	**92.44**	<u>90.26</u>	<u>93.38</u>	**84.13**	**86.27**	**84.96**	**85.40**	**79.58**	**79.36**	**78.24**	**78.86**	**78.41**	**78.26**	**78.01**	**78.68**

shows density-stratified results, where Dynamo maintains strong full-FT BI-RADS AUC and κ across density groups. These gaps suggest multimodal pretraining aligns better with BI-RADS semantics, improving zero-shot and label-efficient transfer.

Generalization to CSAW-CC. Fig. 2(a) indicates that the EMBED-trained Dynamo preserves the lowest calibration errors and strongest 3–5y discrimination on CSAW-CC. At 1–2y, survival-style *Vision-only* baselines (trained on CSAW-CC) can be slightly ahead, but Dynamo consistently outperforms *CLIP-style* models overall, suggesting that its calibration and long-horizon ranking advantages transfer across cohorts despite distribution shift.



(a) All evaluated models (b) Temporal robustness analysis

Fig. 2: **Predictive performance and robustness of Dynamo. (a)** The 1–5-year AUC values and mean AUC for all models on CSAW-CC **(b)Temporal robustness analysis** of the proposed CDE versus a Transformer baseline under extreme resampling and data dropout on EMBED.

Table 3: **Ablation experimental results on EMBED.** We report C-index, 1-Year and 5-Year AUC for risk prediction, and BI-RADS prediction AUC under linear probing (Lin. Probe) and full fine-tuning (Full FT).

Design Choices										Risk Pred. (%)			BI-RADS Pred. (%)	
Loss Ablation			Temporal Encoding		Image Encoder			Multi-Scale		C-index	1Y-AUC	5Y-AUC	Lin. Probe	Full FT
$\mathcal{L}_{\text{long}}$	$\mathcal{L}_{\text{tvqa}}$	$\mathcal{L}_{\text{asym}}$	NeuralCDE	Transformer	Mirai	Mammo		Coarse	Fine					
	✓	✓	✓		✓			✓	✓	72.30	80.26	67.82	61.48	62.51
✓		✓	✓		✓			✓	✓	73.08	82.34	70.88	61.90	63.04
✓	✓		✓		✓			✓	✓	76.91	82.58	71.40	67.24	69.96
✓	✓	✓		✓	✓			✓	✓	75.38	85.24	78.91	70.26	71.93
✓	✓	✓	✓			✓		✓	✓	71.29	74.29	71.93	63.87	64.82
✓	✓	✓	✓		✓			✓		76.41	85.38	80.77	74.48	76.59
✓	✓	✓	✓		✓				✓	<u>80.25</u>	<u>86.51</u>	<u>82.38</u>	<u>76.15</u>	<u>78.45</u>
✓	✓	✓	✓		✓			✓	✓	**82.79**	**91.35**	**85.19**	**78.54**	**79.01**

3.4 Ablation study

We perform comprehensive ablation experiments on EMBED, with results shown in Table 3. The removal of TVQA has the most pronounced effect, producing uniform drops of roughly one order of magnitude across metrics. Replacing the continuous-time NCDE with a temporal Transformer consistently underperforms the full model in the C-index, 1y/5y AUC and in the BI-RADS AUC, highlighting the benefit of continuous-time trajectory modeling. Excluding the asymmetry loss also degrades results, indicating that left–right correspondence is a useful inductive bias. On the image backbone side, Mirai [32] is clearly preferable to a Mammo-only encoder, suggesting that richer pretraining is critical. Finally, multi-scale inputs are complementary: fine-only surpasses coarse-only, and using both scales delivers the best results across all endpoints.

3.5 Temporal Robustness Analysis

Evaluating temporal dynamics, NCDE [16] outperforms a Transformer [28] baseline across a balanced subset of the most regular and irregular screening intervals

(404 patients each), with pronounced gains under irregular sampling (Fig. 2b). Moreover, selectively removing intermediate visits substantially degrades Transformer performance, demonstrating NCDE's robustness to temporal sparsity.

4 Conclusion

In this work, we introduced Dynamo, a multi-modal pretraining framework designed to address key challenges in longitudinal mammography analysis. Our approach remedies the "irregularity gap" of discrete models by employing a Neural Controlled Differential Equation (CDE) to represent irregular patient screening timelines as continuous latent trajectories. Concurrently, it bridges the "temporal reasoning gap" of static vision-language models by introducing a Temporal Visual-Question Answering (TVQA) loss, which compels the model to ground semantic representations in the temporal evolution of imaging data. Evaluations on two large scale longitudinal datasets, EMBED and CSAW-CC, demonstrated the framework's effectiveness. Future work will focus on improving short-term predictive accuracy, validating the approach across more diverse multi-center international cohorts and on datasets that include free-text reports, and extending this continuous-time, query-driven pretraining paradigm to other medical imaging domains characterized by irregular longitudinal data.

Acknowledgments. This work was supported by The Research Council of Norway (Visual Intelligence, grant no. 309439, FRIPRO grants no. 315029 and no. 360068, and IKTPLUSS grant no. 303514).

Disclosure of Interests. The authors declare no competing interests.

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **74**(3), 229–263 (2024)
2. Butcher, J.: Runge-kutta methods. Scholarpedia **2**(9), 3147 (2007)
3. Calvo, M., Montijano, J.I., Randez, L.: A fifth-order interpolant for the dormand and prince runge-kutta method. Journal of computational and applied mathematics **29**(1), 91–100 (1990)
4. Chen, Y., Ren, K., Wang, Y., Fang, Y., Sun, W., Li, D.: Contiformer: continuous-time transformer for irregular time series modeling. In: NeurIPS. Curran Associates Inc. (2023)
5. Donnelly, J., Moffett, L., Barnett, A.J., Trivedi, H., Schwartz, F., Lo, J., Rudin, C.: Asymmirai: Interpretable mammography-based deep learning model for 1–5-year breast cancer risk prediction. Radiology **310**(3), e232780 (2024)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

7. Du, Y., Onofrey, J.A., Dvornek, N.C.: Multi-view and multi-scale alignment for contrastive language-image pre-training in mammography. In: IPMI. Lecture Notes in Computer Science, vol. 15830, pp. 247–262. Springer (2025)
8. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. In: MICCAI. Lecture Notes in Computer Science, vol. 15012, pp. 632–642. Springer (2024)
9. Glenn, W.B., et al.: Verification of forecasts expressed in terms of probability. Monthly weather review **78**(1), 1–3 (1950)
10. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: COLM (2024)
11. Hao, S., Li, S.: A weighted mean absolute error metric for image quality assessment. In: VCIP. pp. 330–333. IEEE (2020)
12. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. Jama **247**(18), 2543–2546 (1982)
13. Hussain, S., Teevno, M.A., Naseem, U., Avalos, D.B.A., Cardona-Huerta, S., Tamez-Peña, J.G.: Multiview multimodal feature fusion for breast cancer classification using deep learning. IEEE Access **13**, 9265–9275 (2025)
14. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilies, G., Smith, G., Woo, M., McAdams, C.R., Newell, M.S., Banerjee, I., Gichoya, J., Trivedi, H.: The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiology: Artificial Intelligence **5**(1), e220047 (2023)
15. Karaman, B.K., Dodelzon, K., Akar, G.B., Sabuncu, M.R.: Longitudinal mammogram risk prediction. In: MICCAI. pp. 437–446. Springer (2024)
16. Kidger, P., Morrill, J., Foster, J., Lyons, T.: Neural controlled differential equations for irregular time series. In: NeurIPS. Curran Associates Inc. (2020)
17. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (Sep 2019)
18. Leung, J.W.T., Sickles, E.A.: Developing asymmetry identified on mammography: Correlation with imaging outcome and pathologic findings. American Journal of Roentgenology **188**(3), 667–675 (2007)
19. Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., Hoi, S.C.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS. pp. 9694–9705 (2021)
20. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012)
21. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019), `https://arxiv.org/abs/1807.03748`
22. Oquab, M., et al.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
24. Sellergren, A., et al.: Medgemma technical report (2025), `https://arxiv.org/abs/2507.05201`
25. Shen, Y., Park, J., Yeung, F., Goldberg, E., Heacock, L., Shamout, F., Geras, K.J.: Leveraging transformers to improve breast cancer classification and risk assessment

with multi-modal and longitudinal data (2023), https://arxiv.org/abs/2311.03217
26. Strand, F.: CSAW-CC (mammography) - a dataset for AI research to improve screening, diagnostics and prognostics of breast cancer (2022)
27. Sun, Z., Thrun, S., Kampffmeyer, M.: Vmra-mar: An asymmetry-aware temporal framework for longitudinal breast cancer risk prediction. In: MICCAI. Lecture Notes in Computer Science, vol. 15974, pp. 660–669. Springer (2025)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. p. 6000–6010. Curran Associates Inc. (2017)
29. Wang, L.: Mammography with deep learning for breast cancer detection. Frontiers in oncology **14**, 1281922 (2024)
30. Wei, X., Tao, Y., Du, C., Zhao, G., Yu, Y., Li, J.: Vikl: A mammography interpretation framework via multimodal aggregation of visual-knowledge-linguistic features (2024), https://arxiv.org/abs/2409.15744
31. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A deep learning mammography-based model for improved breast cancer risk prediction. Radiology **292**(1), 60–66 (2019)
32. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Smith, K., Wan, Y.L., Lamb, L., Hughes, K., Lehman, C., Barzilay, R.: Toward robust mammography-based models for breast cancer risk. Science Translational Medicine **13**(578), eaba4373 (2021)