

Interpretable Concept-Bottleneck Vision-Language Model Predicts Pathologic Complete Response From Longitudinal Breast DCE-MRI

Charlotte N. Li

Montefiore Medical Center and Albert Einstein College of Medicine

Daksh Khurana

Montefiore Medical Center and Albert Einstein College of Medicine

Takouhie Maldjian

Montefiore Medical Center and Albert Einstein College of Medicine

Tim Duong

`tim.duong@einsteinmed.edu`

Montefiore Medical Center and Albert Einstein College of Medicine

Article

Keywords:

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11040164/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Deep learning can predict neoadjuvant treatment response in breast cancer, but most models are black boxes whose reasoning clinicians cannot verify, limiting adoption. We developed an interpretable concept-bottleneck vision-language model (CB-VLM) that predicts pathologic complete response (pCR) from pre-treatment and early-treatment dynamic contrast-enhanced (DCE) MRI, routing every prediction through eleven radiologist-recognizable, BI-RADS-style imaging concepts plus age and molecular subtype. Trained on 753 patients from the multicenter I-SPY2 trial, the model achieved an AUROC of 0.76 on encoder-held-out data ($n = 144$) and generalized to an independent I-SPY1 cohort (AUC 0.71; $n = 110$). Accuracy rose with model confidence (88% versus 58% on borderline predictions), and early change in functional tumor volume was the dominant predictor across both cohorts. To our knowledge, this is the first large vision-language model to generate concept-grounded reasoning for breast DCE-MRI treatment response prediction, reframing a black-box task as transparent, auditable, and generalizable.

Introduction

For more than a decade, deep learning analysis of breast imaging for diagnosis, prognosis, and treatment response prediction advanced primarily through task-specific convolutional neural networks (CNNs) trained on single objectives¹. Although such systems often matched or exceeded expert performance on curated benchmarks, their reliance on large annotated datasets, limited generalizability across scanners and populations, and opacity as black-box classifiers have constrained clinical translation². Performance can degrade under dataset shift (i.e., when models encounter a different dataset) and prospective, multicenter validation remains relatively uncommon. As a result, it is often difficult for clinicians to verify a model's reasoning, anticipate its failure modes, or integrate its output into their own clinical judgment, and questions of accountability for an unexplained recommendation are not easily resolved. Interpretability and transparency, alongside accuracy, are therefore increasingly recognized as important prerequisites for the acceptance and clinical adoption of AI tools³.

Novel model architectures, such as vision-language models and concept bottleneck models, may address this gap. Vision-language models can verbalize their reasoning while processing complex images and text jointly, but hallucination remains a concern because their outputs are routed through many opaque model layers^{4,5}. Concept bottleneck models, by contrast, route their predictions through a layer of clinically grounded concepts, so the basis of each decision is explicit and a clinician can inspect the intermediate concepts on which a prediction rests⁶. Concept bottleneck models have recently been shown to match black-box performance in breast MRI diagnosis while providing intrinsic interpretability⁷. These strengths are complementary: pairing a concept bottleneck with a vision-language model can retain the model's ability to generate a natural language rationale while routing that rationale through a layer of verifiable imaging concepts, directly addressing the hallucination concern.

Building on this complementarity, we combined the two into a concept bottleneck vision-language model (CB-VLM) for longitudinal breast DCE-MRI, developed with fine-tuning and expert-in-the-loop review of

generated reports. The CB-VLM uses two complementary image encoders to produce a shared visual representation, from which it predicts eleven imaging concepts anchored in the standardized ACR BI-RADS lexicon and IBSI-compliant radiomic definitions. These concepts, combined with molecular subtype and age, form an auditable layer from which the language component generates its report, reasoning, and prediction, yielding rationales a radiologist can check against the image.

We evaluated the architecture on prediction of pathologic complete response (pCR) in breast cancer, as pCR is a strong prognostic marker of event-free and overall survival^{8–10}, and residual disease after neoadjuvant therapy identifies patients who may benefit from adjuvant therapy escalation¹¹. Because early treatment imaging is acquired while therapy can still be adapted, development of transparent, checkable and accurate predictors of pCR at this timepoint can have direct therapeutic implications. We evaluated the model on baseline and early treatment DCE-MRI in 753 patients from the multicenter I-SPY2 trial, and tested generalizability on an independent dataset from the multicenter I-SPY1/ACRIN-6657 trial.

Concept bottlenecks have recently been applied to longitudinal volumetric response assessment in neuro-oncology, but without a language component and without external validation¹¹. To our knowledge, this is the first vision-language model to generate concept-grounded reasoning for treatment-response prediction from breast DCE-MRI. It is also the first to couple an auditable concept bottleneck over longitudinal volumetric imaging with a generative language readout, yielding a structured report, an explicit reasoning trace, and a calibrated probability. This interpretable architecture maintains discrimination, transfers to an independent cohort, and communicates its uncertainty on the well-characterized endpoint of pCR.

Results

Study population. Patient Selection flowchart is shown in **Supplementary Fig. 1**. Of 1,104 unique patients identified across the I-SPY2 and ACRIN-6698 TCIA collections, 754 had paired baseline (T0) and early-treatment (T1) DCE-MRI that passed automated preprocessing and tumor segmentation and formed the analytic cohort. Of these, 753 had an evaluable pCR label and were used for all out-of-fold (OOF) discrimination analyses. Patients were partitioned into training, validation, and test sets ($n = 610$, 73 , and 71 , respectively) using a split stratified on treatment arm and molecular subtype. Cohort characteristics, including age, race/ethnicity, menopausal status, molecular subtype, and treatment group, are summarized in **Supplementary Table 1** for I-SPY2, and **Supplementary Table 2** for I-SPY1. An additional 107 patients from the independent I-SPY1/ACRIN-6657 trial were held out as an external validation cohort (see External validation). One training-set patient lacked an evaluable pCR label, leaving 609 training and 753 total patients for the out-of-fold analyses.

Imaging concepts fidelity. The architecture of the CB-VLM is depicted in Fig. 1, and the imaging concepts are described in detail in **Supplementary Table 3**. On the held-out test set ($n = 71$), the dual image encoder recovered continuous imaging concepts with moderate-to-strong fidelity against

algorithmic ground truth (Pearson $r = 0.51-0.76$; Fig. 2a). Sphericity was recovered most faithfully ($r = 0.76$), followed by functional tumor volume (FTV; $r = 0.75$), change in FTV ($r = 0.61$), and change in sphericity ($r = 0.51$). Categorical concepts were predicted with a macro-averaged accuracy of 0.72 (Fig. 2b). Accuracy was highest for mass-versus-non-mass enhancement (0.88), followed by background parenchymal enhancement (0.79), shape (0.76), multifocality (0.73), T1 initial enhancement rate (0.73), shrinkage pattern (0.65), and margin (0.51). Accuracy exceeded the majority-class base rate for five of seven categorical concepts, most clearly for shrinkage pattern, but fell below it for margin and for T1 initial enhancement rate. Overall, size and morphology descriptors were recovered well, whereas longitudinal-change concepts and the most class-imbalanced categorical descriptors were less reliable.

pCR discrimination. The full CB-VLM achieved an AUROC 0.76 on all encoder-held-out patients ($n = 144$) and of 0.66 on the held-out test set ($n = 71$) for pCR prediction (Fig. 3a). To attribute this performance, the model was benchmarked against nested logistic baselines (Fig. 3b). A logistic model using clinical variables alone (molecular subtype and age) achieved an AUROC of 0.636 (95% CI 0.59–0.68). Adding the eleven image-derived concepts improved discrimination to 0.710 (95% CI 0.67–0.75), an increment of + 0.074 (DeLong $P = 4.4 \times 10^{-5}$), indicating that baseline and early-treatment imaging carry predictive signal beyond receptor status and age. Two representative correct, confident cases (a predicted responder and a predicted non-responder), with the concepts that favored versus weighed against pCR, are shown in **Supplementary Fig. 2**.

(a) On the encoder-held-out test set ($n = 144$), the full CB-VLM achieved an AUROC of 0.76 (95% CI 0.68 to 0.84). On the strictly held-out test set ($n = 71$), the CB-VLM achieved an AUROC of 0.66 (95% CI 0.52–0.79). The shaded band shows the bootstrap 95% confidence interval. (b) Incremental value of imaging and VLM architecture in model performance. A logistic regression was used to measure the predictive value of the clinical variables alone (left) versus clinical and imaging variables combined (center) and demonstrated that imaging adds + 0.074 AUC over the clinical model alone. The final CB-VLM architecture added + 0.052 AUROC over the clinical and imaging model (right). Treatment arm was not provided to the model.

Performance by molecular subtype and treatment mechanism groups. Discrimination was measured within molecular-subtype and treatment-mechanism groups (**Supplementary Fig. 3**). Treatment mechanism was used only to define these strata and was not available to the model. Performance was consistent across molecular subtype (AUROC ranged from 0.57–0.68), and was highest in HR+/HER2+ 0.68 ($n = 115$), followed by triple-negative 0.67 ($n = 272$), then HR+/HER2- 0.63 ($n = 310$), and HR-/HER2+ 0.57 ($n = 56$). Across treatment mechanisms, model performance ranged from 0.64 to 0.76. Performance was highest in immunochemotherapy (paclitaxel plus pembrolizumab) (AUROC = 0.76; $n = 59$), followed by other targeted therapies (AUROC 0.73; $n = 343$), HER2-targeted (AUROC = 0.71; $n = 207$), and chemotherapy alone (AUROC 0.64; $n = 144$). Confidence intervals were widest in the smallest treatment group strata, notably HR-/HER2+ ($n = 56$; 95% CI 0.41 to 0.72) and the immunochemotherapy group ($n = 59$; 95% CI 0.63 to 0.87). As a result, these subgroup estimates should be read as exploratory.

Per-variable permutation importance. Per-variable permutation importance was computed within each stratum to identify the variables driving prediction (Fig. 4). Across molecular-subtype panels, early change in FTV was the dominant variable, most prominently in HR+/HER2+ ($\Delta\text{AUC} \approx +0.15$) and triple-negative disease ($\approx +0.07$). Across treatment-mechanism panels, change in FTV again dominated in the HER2-targeted ($\approx +0.07$) and chemotherapy-alone ($\approx +0.06$) groups, with the T1 initial enhancement rate contributing comparably in the chemotherapy-alone group ($\approx +0.06$). By contrast, no single variable showed reliable importance in the immunotherapy group, with confidence intervals spanning zero, consistent with a dissociation between early volumetric shrinkage and immune-mediated response, in which radiographic change may diverge from pathologic response. Permutation importance was computed on the pooled out-of-fold predictions; the ranking is expected to remain stable under the nested re-analysis, though magnitudes may shift.

a.

b.

Model confidence. The expected calibration error (ECE) of the language-model readout was 0.148, with overconfidence at intermediate predicted probabilities (Fig. 5a). Accuracy increased monotonically with model confidence across three bins: 58% for borderline predictions ($|\text{p} - 0.5| < 0.15$; $n = 281$), 70% for moderate-confidence predictions ($0.15 - 0.30$; $n = 234$), and 88% for confident predictions ($|\text{p} - 0.5| \geq 0.30$; $n = 238$) (Fig. 5b). The correlation between confidence and correctness was $r = +0.30$. Restricting predictions to high confidence thus yielded 88% accuracy versus 58% on borderline cases.

External Validation on I-SPY1

The I-SPY2-trained model was applied unchanged to the independent I-SPY1/ACRIN-6657 cohort (Fig. 6). Because I-SPY1 delivered a single standard regimen, stratification by treatment mechanism was not applicable, and treatment arm was not provided to the model in either cohort.

External discrimination and calibration. The language readout achieved an external AUROC of 0.71 (95% CI 0.60 to 0.81), compared with 0.65 for the clinical-only model and 0.72 for the concept-and-clinical logistic teacher (Fig. 6a). The language readout therefore did not retain its internal advantage over the teacher, a reversal of the internal ordering. Calibration was preserved and was numerically better than internally (expected calibration error 0.08 versus 0.15; Fig. 6b). Discrimination by molecular subtype was 0.72 overall (Fig. 6c) but ranged from 0.61 to 0.93 across strata containing as few as 15 patients; these estimates are unstable and are reported for completeness. Within confidence intervals, the model performed similarly on I-SPY1 and I-SPY2, indicating that it is transferrable across cohorts.

External concept fidelity. Concept fidelity was recomputed under scanner and protocol shift using the same fixed thresholds (Fig. 6d). Categorical accuracy was highest for shape (0.83), multifocality (0.77) and background parenchymal enhancement (0.76). Continuous concepts showed reduced fidelity, with sphericity at $r = 0.61$, tumor volume at $r = 0.47$, change in functional tumor volume at $r = 0.39$ and change

in sphericity at $r = 0.22$. The concepts that degraded most are those on which the internal model depended most heavily, offering a concept-level account of the reduced discrimination.

External variable importance. Permutation importance nonetheless placed change in functional tumor volume as the leading variable within every molecular subtype (Fig. 6e), consistent with the internal ranking.

Discussion

To our knowledge, this is the first concept-bottleneck vision-language model applied to breast DCE-MRI, and the first to bring a concept-bottleneck VLM to a treatment response task in any 3D imaging modality. Prior breast-imaging VLMs have addressed mammography-based malignancy classification, BI-RADS density estimation, and report generation, without exposing intermediate reasoning through verifiable imaging concepts^{12–14}. Concept-bottleneck approaches to vision-language modeling have so far been developed for 2D diagnostic image classification, where they match black-box performance while remaining interpretable⁶. Concept bottlenecks have separately been applied to treatment response tasks, such as immunotherapy-response prediction from transcriptomics¹⁵, and longitudinal 3D glioblastoma response on MRI without a language component¹⁶. No concept-bottleneck vision-language model, however, has been applied to longitudinal, volumetric MRI for treatment response prediction, which requires 3D spatial reasoning across timepoints⁶. The present work closes that gap, extending the concept-bottleneck paradigm from 2D diagnostic classification to a longitudinal, volumetric, therapy-response problem. It couples a lesion-focused 3D CNN with a frozen DINOv2 whole-breast encoder to populate an auditable concept layer, which a language model reads out into a report, an explicit reasoning trace, and a calibrated probability.

The central contribution is therefore architectural rather than a gain in raw discrimination. Because every prediction is routed through eleven BI-RADS-style imaging concepts and two clinical variables, it is possible to evaluate which concepts the model measured, how they were weighted, and whether the resulting rationale is clinically coherent, which is a level of verifiability that black-box pCR models do not provide. The model reconstructed clinically meaningful concepts, particularly size and morphology descriptors such as sphericity, FTV, and mass versus non-mass enhancement, with moderate-to-strong fidelity from imaging alone, supporting the feasibility of an interpretable, concept-grounded pipeline in which downstream predictions can be traced to recognizable radiologic features. The weaker recovery of longitudinal change and rare categorical descriptors (e.g., margin, change in sphericity) implies that concept-level transparency and reliability are not yet uniform. As such, these lower fidelity concepts should be read with more caution. Framing this heterogeneity explicitly is itself a benefit of the architecture, because a black-box model offers no comparable self-audit.

Against this interpretable backbone, imaging carried genuine but bounded predictive value. A logistic model using clinical variables alone (molecular subtype and age) achieved an AUROC of 0.636, and adding the eleven image-derived concepts improved discrimination to 0.710, an increment of + 0.074

(DeLong $P = 4.4 \times 10^{-5}$), indicating that baseline and early-treatment imaging carry predictive signal beyond receptor status and age. This is concordant with prior work reporting that pretreatment MRI descriptors such as unifocality and non-spiculated morphology are independently associated with pCR¹⁷. Consistent with recent literature, molecular subtype remained the single strongest predictor, and the incremental imaging contribution, while statistically detectable, was modest; a recent multicenter study likewise found cancer subtype to be the dominant predictor of pCR, with delta-radiomic features contributing a smaller remainder and the combined versus clinical improvement not reaching significance¹⁸.

The held-out AUROC (0.66 on the test set, 0.76 on encoder-held-out patients) is on par with prior DCE-MRI response models, despite relying on earlier imaging timepoints. It compares favorably with the multi-feature MRI logistic models from the I-SPY2 trial, in which combinations of FTV, sphericity, background parenchymal enhancement, and longest diameter outperformed any single feature for pCR prediction¹⁹, and with the addition of apparent diffusion coefficient to FTV, which raised full-cohort AUC to roughly 0.78–0.81 only at mid- and post-treatment timepoints²⁰. Achieving comparable discrimination from only two early timepoints (baseline and after approximately three weeks of therapy) is arguably a more clinically consequential result, because it identifies likely responders and non-responders while treatment course can be modified. Early volumetric response is already being used to guide risk-adapted neoadjuvant strategies: I-SPY2.2 uses the change in functional tumor volume during treatment to redirect predicted non-responders to alternative therapy, while MRI- and PET-guided trials such as TRAIN-3 and PHERGain use early radiological response to de-escalate chemotherapy without compromising outcomes^{21–23}. An interpretable, calibrated predictor available at this early timepoint could support such response-adapted decisions, while allowing clinicians to verify the basis of each prediction.

Early change in FTV was the dominant predictor across nearly all molecular-subtype and treatment-mechanism strata, reinforcing a consistent theme in the neoadjuvant imaging literature. In the ACRIN 6657/I-SPY lineage, FTV measured by DCE-MRI was more accurate than longest diameter or clinical exam for predicting pCR, and early volumetric change carried independent prognostic value^{9,24,25}. The I-SPY2 experience further showed that change in FTV after one cycle predicts both pCR and event-free survival, with a $\geq 65\%$ reduction at 6 weeks predicting a favorable residual cancer burden²¹. That the model independently rediscovered Δ FTV as its most important variable, without being directed to, is biologically plausible and consistent with established response biology, and supports the credibility of the concept bottleneck.

The immunotherapy arm is an instructive exception. No single concept showed reliable importance in the paclitaxel-plus-pembrolizumab group, and confidence intervals spanned zero, consistent with the recognized dissociation between early radiographic shrinkage and pathologic response under immune checkpoint blockade, in which fibrosis, immune infiltration, and tertiary lymphoid structures can occupy the regression bed and mask response on imaging²⁶. This aligns with the broader recognition that volumetric criteria calibrated to cytotoxic and HER2-targeted therapy may not capture immune-related

response kinetics, and argues for immunotherapy-specific concepts or additional biomarkers (for example, ctDNA clearance, which predicted response and outcome in I-SPY2) rather than reliance on early volume change alone²⁷.

Beyond transparency, the calibrated confidence of the readout is itself clinically actionable. Accuracy rose from 58% on borderline predictions to 88% on confident predictions, so the model can flag the cases it is likely to get right and defer the rest; in a triage framing, a reader can concentrate scrutiny on uncertain cases rather than treating all predictions as equally reliable. Chain-of-thought decoding did not improve discrimination ($\Delta\text{AUC} = 0.014$), but it produced an explicit, auditable reasoning trace alongside a calibrated readout, consistent with evidence that structured reasoning improves interpretability and calibration more than raw accuracy in medical language models^{28,29}.

Generalizability was tested directly. Applied unchanged to I-SPY1, a trial conducted more than a decade before I-SPY2 with different scanners and acquisition protocols, the language readout achieved an AUROC of 0.71 with an expected calibration error of 0.077, better than its internal calibration. That the model transferred was enabled by withholding treatment arm during development, since I-SPY1 delivered a single standard regimen to which no I-SPY2 arm corresponds. Just as informatively, the concepts that degraded most under scanner and protocol shift were those on which the internal model depended most heavily, so the concept layer offers a mechanistic, concept-level account of the reduced external discrimination rather than an unexplained performance drop.

Two features distinguish this work methodologically from prior work. First, interpretability is built in rather than retrofitted: clinically meaningful concepts are predicted before the outcome, and the concept-fidelity analysis makes the model's own limitations transparent, bounding the trust that should be placed in explanations invoking weaker concepts. Second, the calibration analysis translates directly into clinical utility, suggesting a triage strategy in which confident model outputs inform early treatment adaptation while uncertain cases default to standard serial imaging and biopsy. Several limitations apply. The analysis is confined to the I-SPY2/ACRIN-6698 population with one external cohort I-SPY1; although these are both multi-institutional datasets, testing on additional data would improve generalizability. Second, certain concepts were derived more reliably than others: background parenchymal enhancement was approximated from cropped ROIs because full-breast segmentations were unavailable, and class imbalance degraded rare categorical concepts (e.g., multifocality, margin), where the model defaulted to the majority class. Third, the held-out test set is small ($n = 71$), widening confidence intervals in the smallest strata, and the pipeline was trained against algorithmically derived concept ground truth rather than radiologist-adjudicated labels, so agreement with expert readers remains to be established. Finally, the endpoint is pCR rather than long-term event-free or distant recurrence-free survival; prospective evaluation linking model-guided adaptation to oncologic outcomes remains the essential next step.

The concept-bottleneck vision-language architecture presented here shows that treatment response prediction from early longitudinal breast DCE-MRI can be made fully auditable by routing predictions

through BI-RADS-grounded imaging concepts, calibrated confidences, and an explicit chain of thought. The strict bottleneck preserved discrimination while enabling granular attribution of signal sources, revealing molecular subtype as the single strongest variable and imaging concepts as a significant but smaller increment. These findings suggest that the near-term value of interpretable VLMs in response prediction may lie less in improving discrimination than in providing the transparency needed to understand what drives a prediction, to identify when imaging adds genuine value, and to build the clinical trust required for prospective evaluation.

Methods

Data sources. This retrospective study analyzed longitudinal breast DCE-MRI from the multicenter I-SPY1 and I-SPY2 trials.^{8–10} Imaging for both cohorts was downloaded from The Cancer Imaging Archive. For I-SPY1, complete pre-treatment (T0) and early post-chemotherapy (T1) imaging was obtained from the ACRIN-6657 trial. To create this cohort, 129 patients with complete T0 and T1 imaging were downloaded. Of the 129 patients, 115 had a well-segmented tumor at T0 and T1, 110 had available pCR data, and 107 had complete clinical data available. For I-SPY2, imaging was acquired from the main I-SPY2 collection and ACRIN-6698 diffusion-weighted substudy. To create this cohort, these collections were merged and deduplicated by subject identifier, yielding 1,104 unique patients, of which 350 were excluded during automated preprocessing and tumor segmentation, leaving 754 patients with paired T0 and T1 imaging (Supplementary Fig. 1). One patient was excluded due to missing clinical data. Out-of-fold analyses were performed in the 753 patients with an evaluable pCR label. Clinical data (pCR, hormone receptor and HER2 status, age) for both trials were drawn from BreastDCEDL, a curated compilation of I-SPY1, I-SPY2 and ACRIN trial data¹¹. I-SPY1 was held out entirely as an external validation cohort and was not used at any stage of model development. I-SPY2 patients were split into training, validation and test sets.

Image preprocessing. Multi-phase DCE volumes were extracted from raw DICOM files. For each patient timepoint, pre-contrast, early post-contrast, and late post-contrast kinetics volumes were retained, having been shown to be the most informative for outcome prediction¹¹. Volumes were then standardized, and a region of interest (ROI) centered on the tumor centroid was extracted for model training.

Imaging concepts. The model predicts eleven imaging concepts (Table 2). Seven capture baseline tumor morphology (functional tumor volume or FTV, sphericity, shape, margin, mass versus non-mass enhancement, multifocality, and background parenchymal enhancement) and were derived using BI-RADS grounded, IBSI-compliant methodology^{30,31}. One kinetics concept, wash-in enhancement rate at T1, captures early enhancement dynamics, a feature previously shown to predict prognosis³². The remaining three concepts (change in FTV, change in sphericity, and shrinkage pattern) describe delta radiomics between T0 and T1. **Supplementary Table 3** gives, for each concept, the operational definition, the variable type, the feature or rule from which the reference value was computed, and, for categorical

concepts, the cut point. Categorical thresholds were calibrated on the I-SPY2 training set and held fixed for all subsequent evaluation, including the external cohort.

Image encoders. The CB-VLM consisted of two sequentially trained parts: a) a dual image encoder trained to predict the eleven imaging concepts, and (b) an instruction-tuned language model that receives the predicted concepts alongside clinical variables, and generates a radiology semi-structured report, chain of thought and pCR prediction (Fig. 1).

To capture fine-grained tumor detail in the context of the full breast, two image encoders, one for lesion-specific attention and one for full-breast context, were created alongside eleven concept multilayer perceptron (MLP) heads, trained jointly using a weighted multitask loss. The first encoder, a 3D residual CNN, was trained over both timepoints to capture fine-grained tumor detail. The CNN first learned a 3D feature representation of the lesion from the cropped tumor region of interest (ROI). This feature map was pooled into a single lesion vector using a learned spatial attention map as a weighted average. The CNN was supervised by a soft-Dice loss, which penalized the model for lesion vector attention that was outside of the ground truth tumor segmentation, or failed to cover it. To ensure the feature maps were regularized over the entire lesion, an auxiliary tumor segmentation head was added to the first encoder and used to decode and correct the 3D feature representations using a binary cross-entropy (BCE) and soft-Dice loss^{33,34} The second encoder, a frozen DINOv2 vision transformer was used to produce full-breast representations by extracting per-patient feature vectors from whole-breast axial slices⁵.

Concept bottleneck. To construct the interpretable model concept bottleneck, the two encoder outputs were projected, concatenated, and then reduced into 1x256 imaging vectors by a small fusion multilayer perceptron (MLP) head. For each patient, the T0 and T1 imaging vectors were then joined into a longitudinal representation, consisting of T0, T1 and the difference between the two vectors. These longitudinal representations were then fed into eleven MLP heads, each of which was trained to predict an imaging concept. MSE loss was used for each continuous concept, and CE loss was used for each categorical variable. The 3D CNN, fusion MLP, and eleven per-concept MLPs were trained jointly using a weighted multitask loss, combining the soft-Dice, BCE, MSE, and CE losses from each training objective. The DINOv2 encoder remained frozen throughout the process, used only to generate full-breast representations.

Language model. The eleven predicted imaging concepts were concatenated with molecular subtype and age from the I-SPY2 cohort to form interpretable concept vectors. To generate model reasoning and predictions, the concept bottleneck vectors were mapped by a small projector into the token-embedding space of an instruction-tuned language model (Qwen2.5) as a set of soft visual tokens³⁵. These tokens were prepended to a tokenized text prompt consisting of concept values, clinical context and the model question³⁶. Conditioned on these tokens, the LLM was trained to generate a structured BI-RADS-style report and an explicit chain of thought. The model was also trained to produce the pCR prediction as a percentage likelihood generated from normalized readout (percentage pCR / (percentage pCR + percentage no pCR)).

The model was fine-tuned with 4-bit quantized low-rank adaptation (QLoRA) and supervised by soft-label knowledge distillation, updating the projector layers and low-rank adapters while keeping the base model and vision encoders frozen³⁷.

QLoRA training was supervised through soft-label knowledge distillation³⁸. A logistic regression teacher model was trained on the observed molecular subtype and age, along with the eleven image-derived concepts (one-hot-encoded categoricals and standardized continuous measures). Per-patient pCR percentages were generated using five-fold out-of-fold cross-validation. The language model was then trained to match these generated probabilities using a two-way soft CE loss.

To supervise the model's reasoning, a structured chain-of-thought target was constructed for each patient, based on the teacher model's decision basis. The teacher model assigned each of the imaging and clinical variables a signed contribution score, and the weight was multiplied by the patient's standardized values for continuous concepts. The four most contributive concepts were written into the chain of thought, teaching the LLM to reproduce the teacher's decision basis.

Concept Fidelity. We evaluated how accurately the model predicted each imaging concept by comparing its predictions to the algorithmically derived ground truth. Pearson r and mean absolute error (MAE) were used to measure the fidelity of continuous concepts, and percent accuracy was used for categorical concepts. Categorical thresholds were held fixed at the I-SPY2 calibrated cut points.

pCR discrimination and calibration. Receiver-operator curves (ROC) were generated, and area under the curve (AUROC) was reported with 95% confidence intervals obtained by patient-level bootstrap resampling over ($n = 2000$) resamples³⁹. The model was provided with patient molecular subtype and age alongside baseline (T0) and early treatment (T1) cropped and full-breast DCE-MRI. Treatment arm was withheld from the model at both training and inference, and no arm indicator entered the concept bottleneck, the language prompt, or the distillation teacher. This was a deliberate design choice, as I-SPY1 delivered a single standard regimen, so a model that depended on arm assignment could not have been applied to the external cohort without modification. Calibration of the model's language readout was summarized by the expected calibration error⁴⁰.

Attribution and interpretability. To determine which variables drove the pCR prediction, per-variable permutation importance was computed using permutation feature importance (shuffling each variable across patients and measuring the resulting change in AUC). This analysis was performed at both the concept-bottleneck level and within the language model's text prompt. To further attribute predictive signal between imaging and clinical channels, imaging concepts were ablated against the molecular subtype and age using logistic-regression probes over each input subset. Tumor-grounded fidelity was quantified by intersection-over-union of the CNN attention map against the lesion mask and through deletion/insertion faithfulness⁴¹. Further clinical analyses were done by splitting the I-SPY2 cohort into four treatment mechanism groups (immunochemotherapy, chemotherapy alone, HER2-targeted, and

other targeted therapies) and measuring the model's out-of-fold performance in each. Model performance was also measured across the different molecular subtypes.

External Validation Cohort

The I-SPY1/ACRIN-6657 cohort was reserved as an independent external test set. Analytic cohorts differed by the variables each analysis required. All 115 quality-controlled subjects with paired T0 + T1 imaging were used for concept fidelity. The 110 patients with an available pathologic complete response entered the 7B language-readout evaluation; and the 107 with complete pathologic complete response, molecular subtype, and age entered the concept-and-clinical and subtype-stratified analyses. The model was frozen after I-SPY2 development and applied unchanged: no weights, thresholds, concept cut points or calibration parameters were refitted on I-SPY1 data. I-SPY1 preceded I-SPY2 and differs from it in scanner generation, acquisition protocol and treatment era, and all patients received a single standard anthracycline and taxane regimen, so treatment-arm stratification is not applicable to this cohort. External evaluation comprised three analyses. First, discrimination for pCR was compared across the clinical-only logistic model, the concept-and-clinical logistic teacher, and the language readout, with AUROC and 95% bootstrap confidence intervals computed as for the internal cohort. Second, calibration of the language readout was summarized by expected calibration error and displayed as a reliability diagram. Third, concept fidelity was recomputed against algorithmically derived reference values using the same procedure and the same fixed thresholds as in I-SPY2, and per-variable permutation importance was recomputed within each molecular subtype to test whether the internal ranking of concepts replicated under scanner and protocol shift. Analyses were performed in Python 3.11 (NumPy 2.4, pandas 3.0, SciPy 1.17, scikit-learn 1.9, scikit-image 0.26, SimpleITK 2.5, Matplotlib 3.10); the 7B language readout was fine-tuned on an NVIDIA A100 GPU (Google Colab) with PyTorch, Transformers, and PEFT (4-bit QLoRA).

Statistics and reproducibility. Discrimination was summarized by the AUROC curve with 95% confidence intervals from patient-level bootstrap resampling over 2,000 resamples³⁹. Paired comparisons between nested models on the same patients were tested with the two-sided DeLong test for correlated ROC curves; alpha was set at 0.05 and exact P values are reported. Continuous concept fidelity was summarized by Pearson correlation with Fisher-transformed 95% confidence intervals and by mean absolute error; categorical concept fidelity was summarized by classification accuracy with Wilson 95% confidence intervals, referenced against the majority-class base rate.

Ethics approval.

This retrospective analysis used publicly available, de-identified imaging and clinical data from the I-SPY1, I-SPY2 and ACRIN-6657/6698 collections distributed through The Cancer Imaging Archive. Datasets were de-identified prior to public release in accordance with the HIPAA Safe Harbor method. Under the U.S. Common Rule (45 CFR 46.104(d)(4)), secondary research using publicly available information recorded such that subjects cannot be identified does not constitute human subjects

research and is exempt from Institutional Review Board (IRB) review. The original trials obtained informed consent and IRB approval at each participating site.

Use of large language models. Qwen2.5-7B is a component of the model under study and its role is described above. Separately, during preparation of this manuscript, OpenEvidence was used to help identify and summarize published literature relevant to the Introduction and Discussion. Claude (Anthropic) was used to edit grammar, revise/restructure sections of the drafted manuscript, and generate initial figures that were subsequently revised for publication. All LLM-assisted output was reviewed and edited by the authors, and the authors take full responsibility for the accuracy and integrity of the manuscript.

Declarations

Data availability

The imaging data analyzed in this study are publicly available from The Cancer Imaging Archive (<https://www.cancerimagingarchive.net/access-data/>), which contains the I-SPY2 collection, the ACRIN-6698 collection, and the ACRIN-6657/I-SPY1 collection⁴². Clinical variables were obtained from BreastDCEDL¹¹ (<https://github.com/naomifridman/BreastDCEDL>).

Code Availability. The code for model training and MRI feature generation is publicly available on GitHub under an open-source MIT license at <https://github.com/pikanaeri/Breast-MRI-Features>.

Acknowledgements. The authors thank the investigators and participants of the I-SPY1, I-SPY2 and ACRIN trials, and The Cancer Imaging Archive, for making these data publicly available. This study received no funding.

Author contributions. C.N.L. designed the model architecture, implemented the training and evaluation pipeline, performed the analyses and drafted the manuscript. D.K. contributed to data curation, preprocessing and concept derivation, and reviewed the analyses. T.M. provided radiologic expertise, reviewed the imaging concept definitions and the generated reports, and revised the manuscript for clinical content. T.Q.D. supervised the study and revised the manuscript. All authors reviewed and approved the final manuscript.

Competing interests. All authors declare no financial or non-financial competing interests.

Ethics approval and consent to participate. This study analyzed publicly available, fully de-identified data and does not constitute human subjects research under 45 CFR 46.104(d)(4); Institutional Review Board review and individual patient consent were therefore not required. Full detail is given in Methods. The original I-SPY1, I-SPY2 and ACRIN trials obtained informed consent and IRB approval at each participating site.

References

1. Reis, T.C. Deep learning in oncology: Transforming cancer diagnosis, prognosis, and treatment. *Emerging Trends in Drugs, Addictions, and Health* 5, 100171 (2025).
2. Kuchana, S.K., et al. Artificial Intelligence in Oncology: A Comprehensive Cross-Cancer Translational Readiness Analysis Across 18 Malignancies. *Cancers (Basel)* 18(2026).
3. Alum, E.U., et al. Overcoming the Black Box Challenge: Building Trust in Artificial Intelligence Algorithms in Oncology. *Technol Cancer Res Treat* 25, 15330338261434649 (2026).
4. Dutta, N., Bose, K., Syailendra, E., Chu, L. & Gupta, P. Vision-language models in diagnostic imaging: review of technical advances, clinical validation, and practical deployment. *Int J Med Inform* 208, 106227 (2026).
5. Oquab, M., et al. DINOv2: Learning Robust Visual Features without Supervision. (2023).
6. Patricio, C., Rio-Torto, I., Cardoso, J.S., Teixeira, L.F. & Neves, J.C. CBVLM: Training-free explainable concept-based Large Vision Language Models for medical image classification. *Computers in Biology and Medicine* 198, 111145 (2025).
7. Qu, J., et al. Multiparameter concept-based interpretable model for early breast cancer diagnosis and structured reporting: a multi-center, multi-reader, radiologist-in-the-loop study. *BMC Medicine* 24(2026).
8. Wang, H. & Yee, D. I-SPY 2: a Neoadjuvant Adaptive Clinical Trial Designed to Improve Outcomes in High-Risk Breast Cancer. *Current Breast Cancer Reports* 11, 303–310 (2019).
9. Hylton, N.M., et al. Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy—results from ACRIN 6657/I-SPY TRIAL. *Radiology* 263, 663–672 (2012).
10. Hylton, N.M., et al. Neoadjuvant Chemotherapy for Breast Cancer: Functional Tumor Volume by MR Imaging Predicts Recurrence-free Survival-Results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. *Radiology* 279, 44–55 (2016).
11. Fridman, N., Solway, B., Fridman, T., Barnea, I. & Goldstein, A. BreastDCEDL: A standardized deep learning-ready breast DCE-MRI dataset of 2,070 patients. *Scientific Data* 13, 264 (2026).
12. Chen, X., et al. Leveraging pretrained vision-language model for enhanced breast cancer diagnosis with multi-view mammography. *Medical Physics* 53, e70261 (2026).
13. Zhu, J., Huang, F., Luo, Q. & Chen, H. A Benchmark for Breast Cancer Screening and Diagnosis in Mammogram Visual Question Answering. *Nature Communications* 16, 11683 (2025).
14. Vo, H.Q., et al. Frozen Large-Scale Pretrained Vision-Language Models are the Effective Foundational Backbone for Multimodal Breast Cancer Prediction. *IEEE Journal of Biomedical and Health Informatics* 29, 3234–3246 (2025).
15. Shen, W., et al. Generalizable AI predicts immunotherapy outcomes across cancers and treatments. *Nat Med* 32, 3010–3022 (2026).
16. Tarek, A., et al. TRACE: A Concept Bottleneck Model for Longitudinal 3D Glioblastoma Response Assessment. *arXiv* (2026).

17. Malhaire, C., et al. Exploring the added value of pretherapeutic MR descriptors in predicting breast cancer pathologic complete response to neoadjuvant chemotherapy. *European Radiology* 33, 8142–8154 (2023).
18. Li, H., et al. Mid-Treatment Delta MRI Radiomics Enhances Sensitivity and Risk Stratification for Pathological Complete Response in Breast Cancer: A Multicenter Study. *Academic Radiology* 33, 3850–3861 (2026).
19. Li, W., et al. Predicting breast cancer response to neoadjuvant treatment using multi-feature MRI: results from the I-SPY 2 TRIAL. *NPJ Breast Cancer* 6, 63 (2020).
20. Li, W., et al. Additive value of diffusion-weighted MRI in the I-SPY 2 TRIAL. *J Magn Reson Imaging* 50, 1742–1753 (2019).
21. Louis, F.M., et al. MRI-based personalisation of neoadjuvant chemotherapy duration in HER2-positive early breast cancer (TRAIN-3): primary results from a multicentre, single-arm, phase 2 study. *Lancet Oncol* 27, 349–360 (2026).
22. Perez-Garcia, J.M., et al. 3-year invasive disease-free survival with chemotherapy de-escalation using an (18)F-FDG-PET-based, pathological complete response-adapted strategy in HER2-positive early breast cancer (PHERGain): a randomised, open-label, phase 2 trial. *Lancet* 403, 1649–1659 (2024).
23. Shatsky, R.A., et al. Datopotamab-deruxtecan plus durvalumab in early-stage breast cancer: the sequential multiple assignment randomized I-SPY2.2 phase 2 trial. *Nat Med* 30, 3737–3747 (2024).
24. Guirguis, M.S., et al. Early tumor volume reduction by breast DCE MRI predicts pathologic complete response to neoadjuvant therapy in triple negative breast cancer. *NPJ Breast Cancer* 11, 129 (2025).
25. Panthi, B., et al. Assessment of Response to Neoadjuvant Systemic Treatment in Triple-Negative Breast Cancer Using Functional Tumor Volumes from Longitudinal Dynamic Contrast-Enhanced MRI. *Cancers (Basel)* 15(2023).
26. Topalian, S.L., et al. Neoadjuvant immune checkpoint blockade: A window of opportunity to advance cancer immunotherapy. *Cancer Cell* 41, 1551–1566 (2023).
27. Magbanua, M.J.M., et al. Circulating tumor DNA refines risk stratification of neoadjuvant therapy-resistant breast tumors. *Nat Commun* 16, 9945 (2025).
28. Wang, X., et al. Reasoning-driven large language models in medicine: opportunities, challenges, and the road ahead. *The Lancet Digital Health* 8, 100931 (2026).
29. Shah, H.A. & Househ, M. Chain of Thought Strategy for Smaller LLMs for Medical Reasoning. *Studies in Health Technology and Informatics* 327, 783–787 (2025).
30. American College of, R. *Breast Imaging Reporting and Data System (BI-RADS) Atlas*, (American College of Radiology, Reston, VA, 2013).
31. Zwanenburg, A., Vallieres, M. & Abdalah, M.A. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* 295, 328–338 (2020).

32. Li, X., et al. DCE-MRI analysis methods for predicting the response of breast cancer to neoadjuvant chemotherapy: pilot study findings. *Magnetic Resonance in Medicine* 71, 1592–1602 (2014).
33. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. in 2016 *Fourth International Conference on 3D Vision (3DV)* 565–571 (2016).
34. Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z. & Tu, Z. Deeply-Supervised Nets. in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Vol. 38 562–570 (PMLR, 2015).
35. Qwen, et al. Qwen2.5 Technical Report. (arXiv, 2024).
36. Liu, H., Li, C., Wu, Q. & Lee, Y.J. Visual Instruction Tuning. in *Advances in Neural Information Processing Systems* 36 34892–34916 (Neural Information Processing Systems Foundation, Inc., 2023).
37. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. in *Advances in Neural Information Processing Systems* 36, Vol. 36 10088–10115 (Neural Information Processing Systems Foundation, Inc., 2023).
38. Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. (arXiv, 2015).
39. Efron, B. & Tibshirani, R.J. *An Introduction to the Bootstrap*, (Chapman & Hall/CRC, Boca Raton, FL, 1994).
40. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K.Q. On Calibration of Modern Neural Networks. in *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 (eds. Doina, P. & Yee Whye, T.) 1321–1330 (PMLR, Proceedings of Machine Learning Research, 2017).
41. Selvaraju, R.R., et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. in *Proceedings of the IEEE international conference on computer vision* 618–626 (2017).
42. Prior, F., et al. The public cancer radiology imaging collections of The Cancer Imaging Archive. *Sci Data* 4, 170124 (2017).

Figures

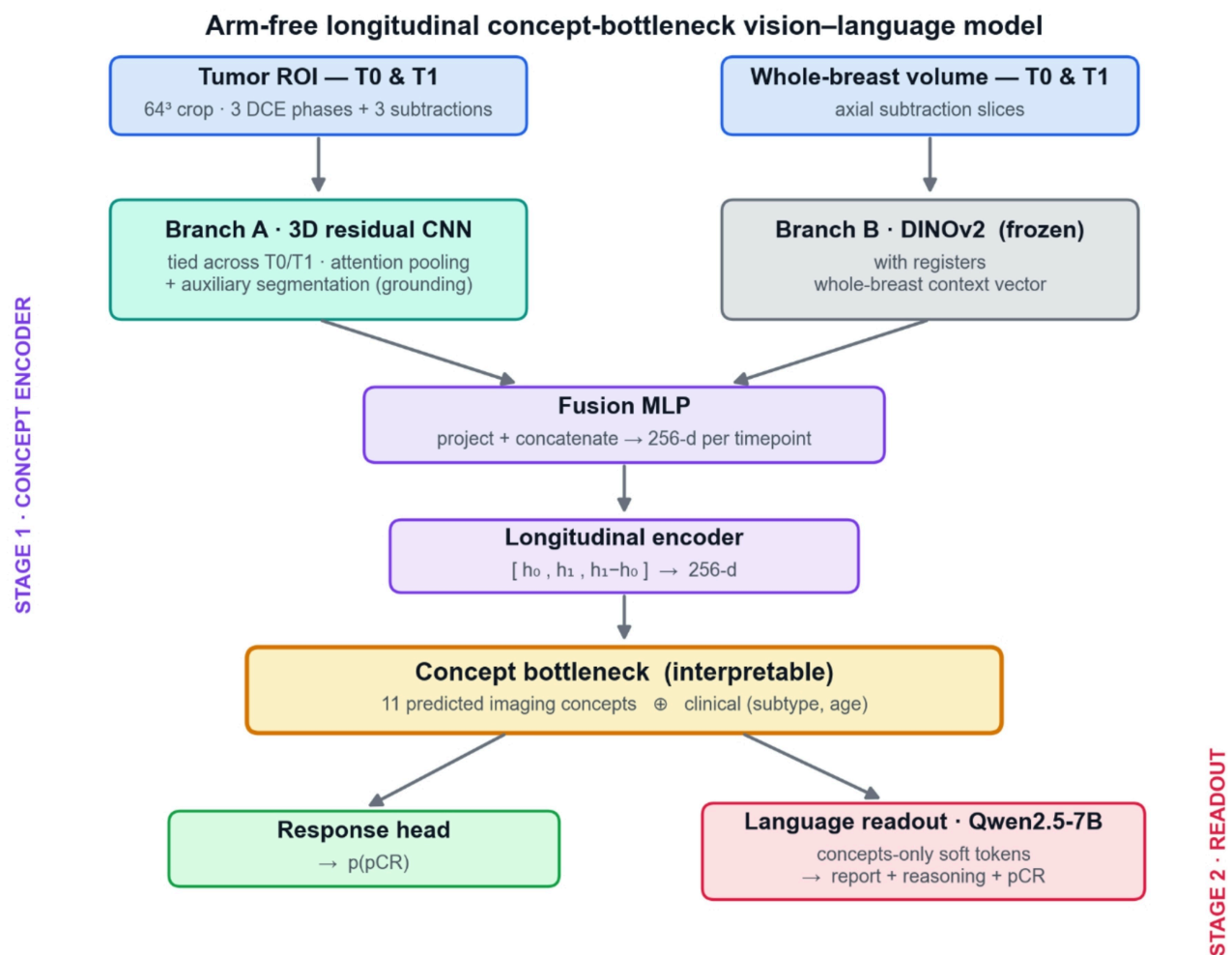


Figure 1

Architecture of the concept-bottleneck vision-language model. The model consists of two parts: an imaging concept encoder, and an interpretable natural language readout. The imaging encoder consists of two models: a lesion-specific CNN (Branch A) and a frozen VLM encoder (Branch B), trained to capture the whole-breast volume. These two encoders feed into eleven MLP, each of which predicts a concept. These eleven concepts, combined with clinical variables form the interpretable concept bottleneck, which is then used to predict patient pCR and a per-patient report.

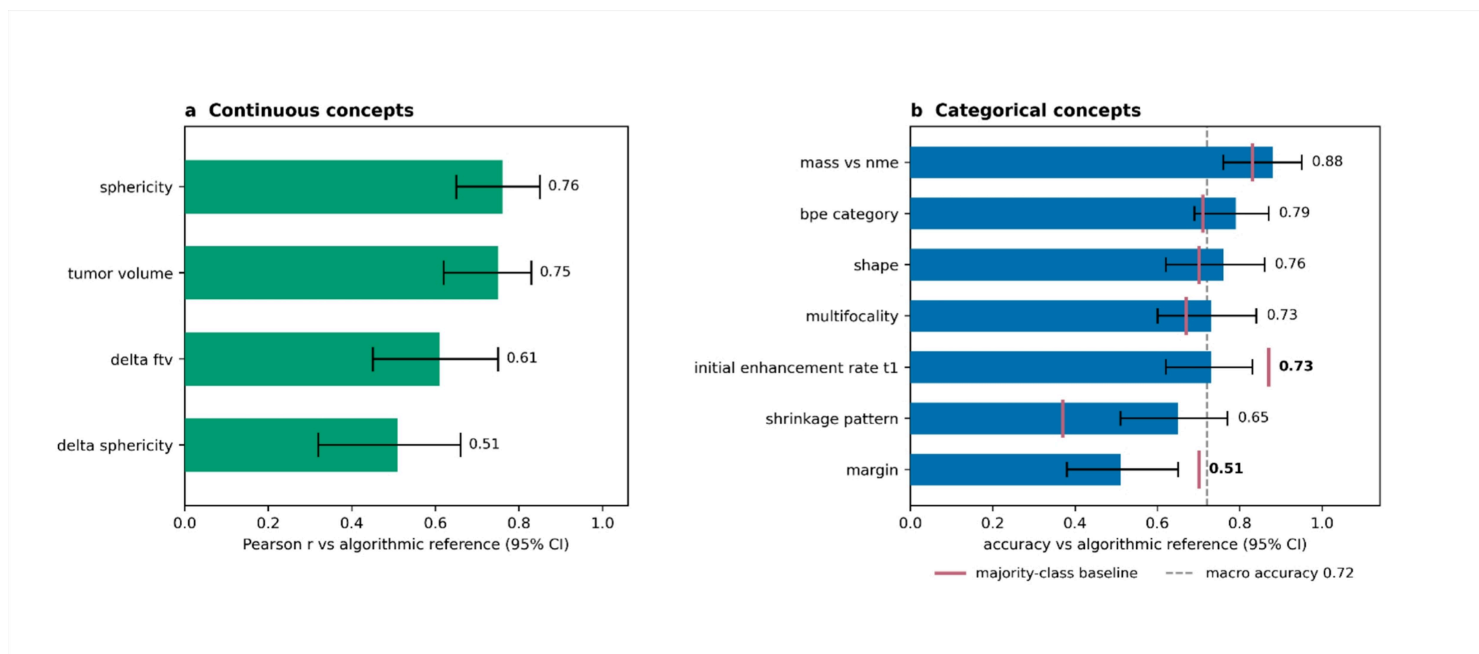


Figure 2

Imaging concepts are recovered with moderate-to-strong fidelity. Concept predictions on the held-out I-SPY2 test set ($n = 71$) were compared against algorithmically derived reference values. a, Continuous concepts were assessed by Pearson r with Fisher 95% confidence intervals. Sphericity (0.76) and functional tumor volume (0.75) were recovered most faithfully, and the longitudinal-change concepts least (change in functional tumor volume 0.61, change in sphericity 0.51). b. Categorical concepts were assessed by percent classification accuracy with Wilson 95% confidence intervals. Red markers denote each concept's majority-class base rate, and the grey dashed line marks the macro-averaged accuracy of 0.72. Accuracy exceeded the majority-class base rate for five of seven concepts, most clearly for shrinkage pattern, but fell below it for margin and for T1 initial enhancement rate (bold labels). Size and morphology descriptors were recovered well, whereas longitudinal change concepts and those with the most imbalanced class distributions were less reliable.

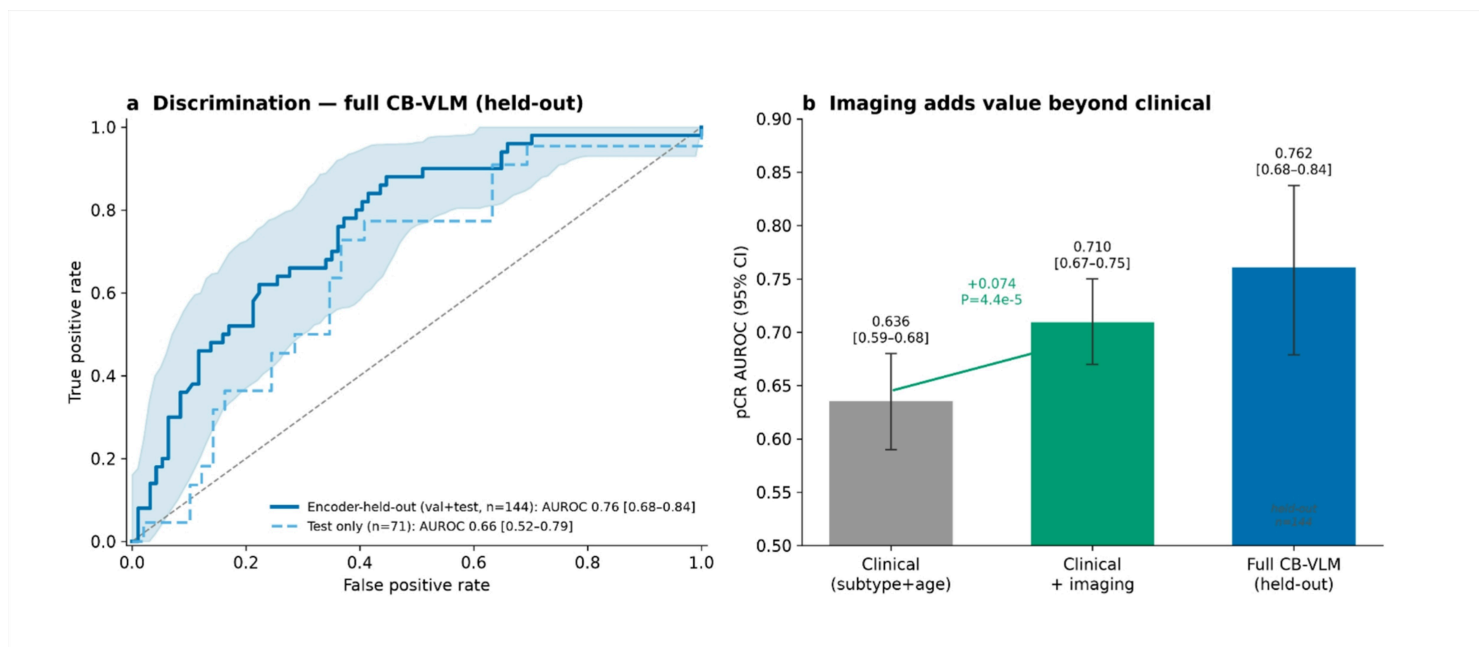
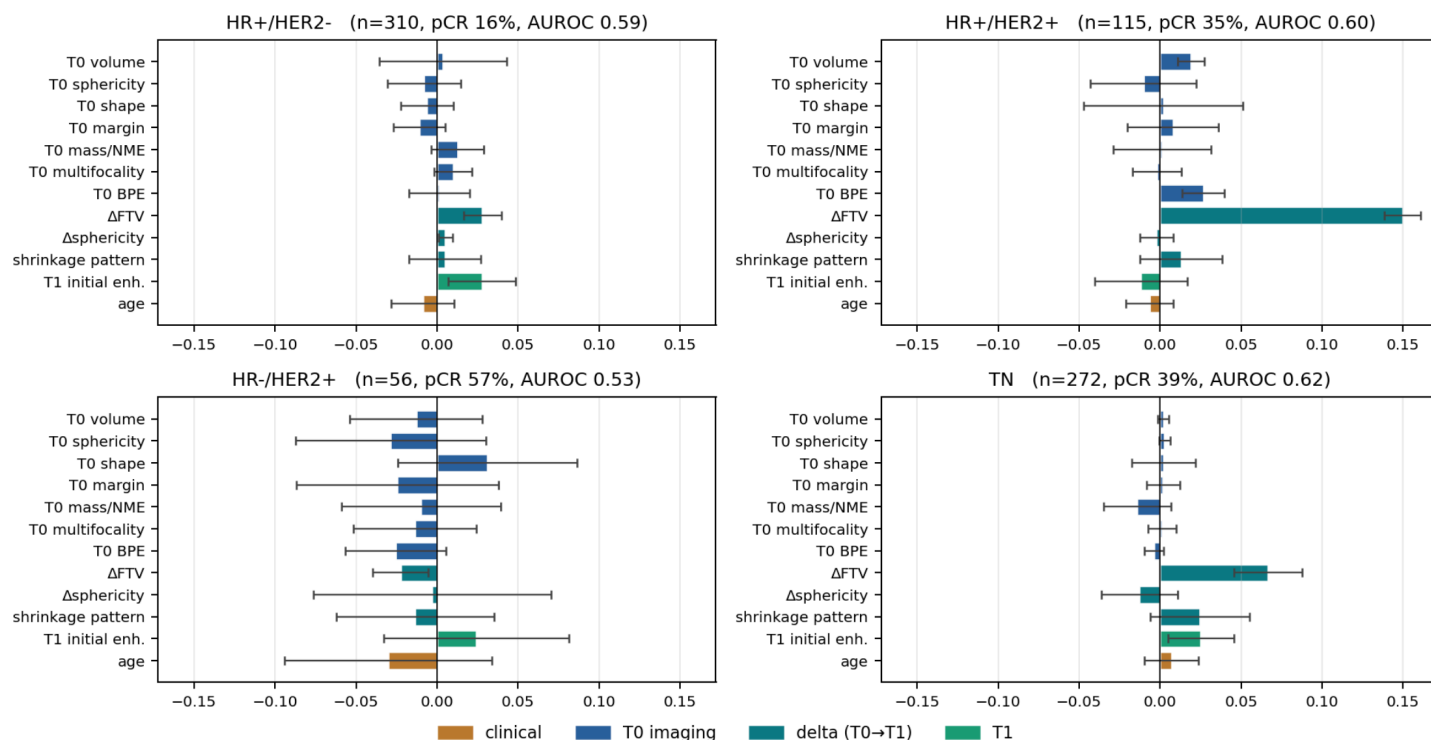


Figure 3

Imaging concepts add discrimination beyond clinical variables.

(a) On the encoder-held-out test set (n=144), the full CB-VLM achieved an AUROC of 0.76 (95% CI 0.68 to 0.84). On the strictly held-out test set (n=71), the CB-VLM achieved an AUROC of 0.66 (95% CI 0.52-0.79). The shaded band shows the bootstrap 95% confidence interval. (b) Incremental value of imaging and VLM architecture in model performance. A logistic regression was used to measure the predictive value of the clinical variables alone (left) versus clinical and imaging variables combined (center) and demonstrated that imaging adds +0.074 AUC over the clinical model alone. The final CB-VLM architecture added +0.052 AUROC over the clinical and imaging model (right). Treatment arm was not provided to the model.

a.



b.

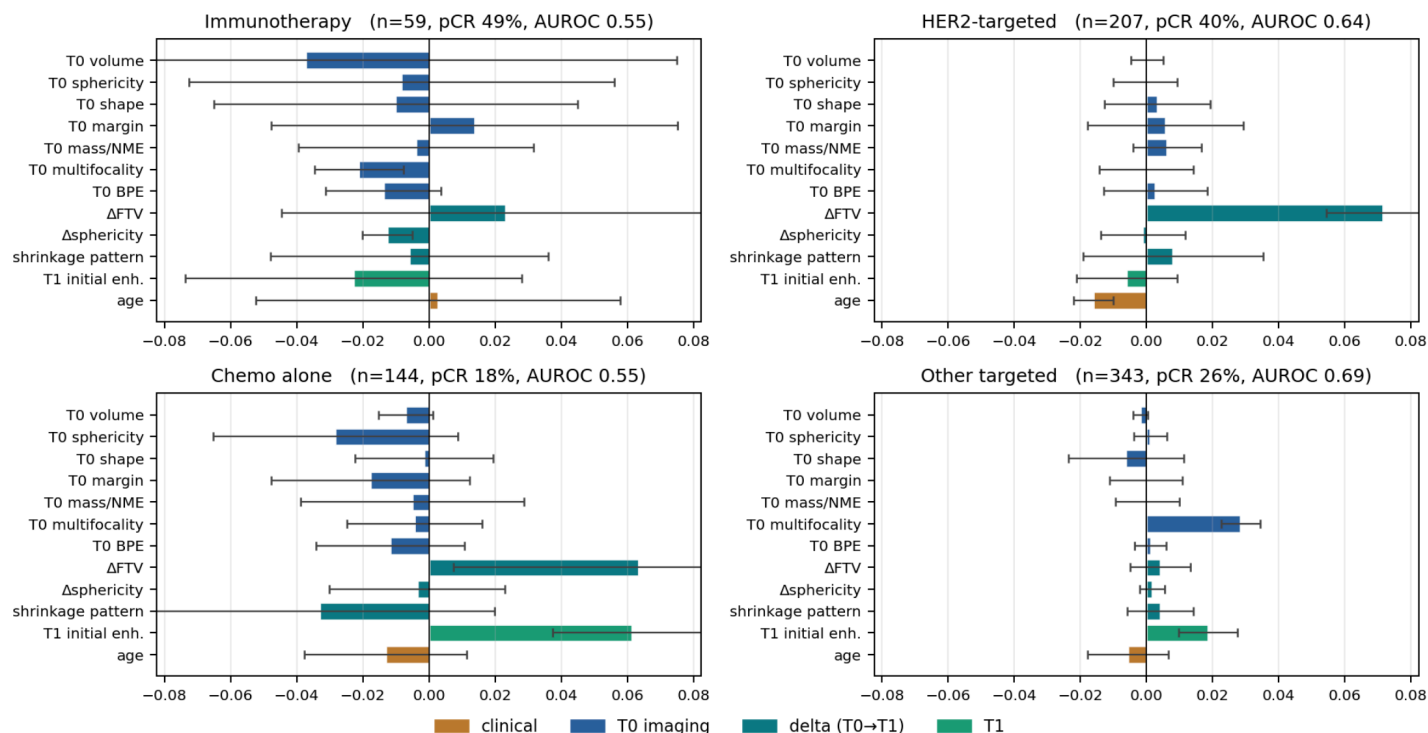


Figure 4

Early change in functional tumor volume dominates prediction across subtypes. Per-variable permutation importance was measured as the change in AUROC when a variable was shuffled across patients; error bars represent ± 1.96 SD over permutation repeats. Panels show molecular-subtype strata (a) and treatment-mechanism strata (b); the sample size and pCR rate for each stratum are shown in the panel headers. Change in functional tumor volume was the leading variable in nearly all strata; no single

variable showed reliable importance in the immunotherapy group, where confidence intervals spanned zero. Within-stratum AUROC (panel headers) reflects discrimination with molecular subtype held constant and is therefore lower than the full-model, subtype-inclusive values in Supplementary Figure 3; the reduction reflects the dominant contribution of subtype. Total n = 753. Panel-header AUROCs are within-stratum, cross-validated logistic estimates over the ground-truth.

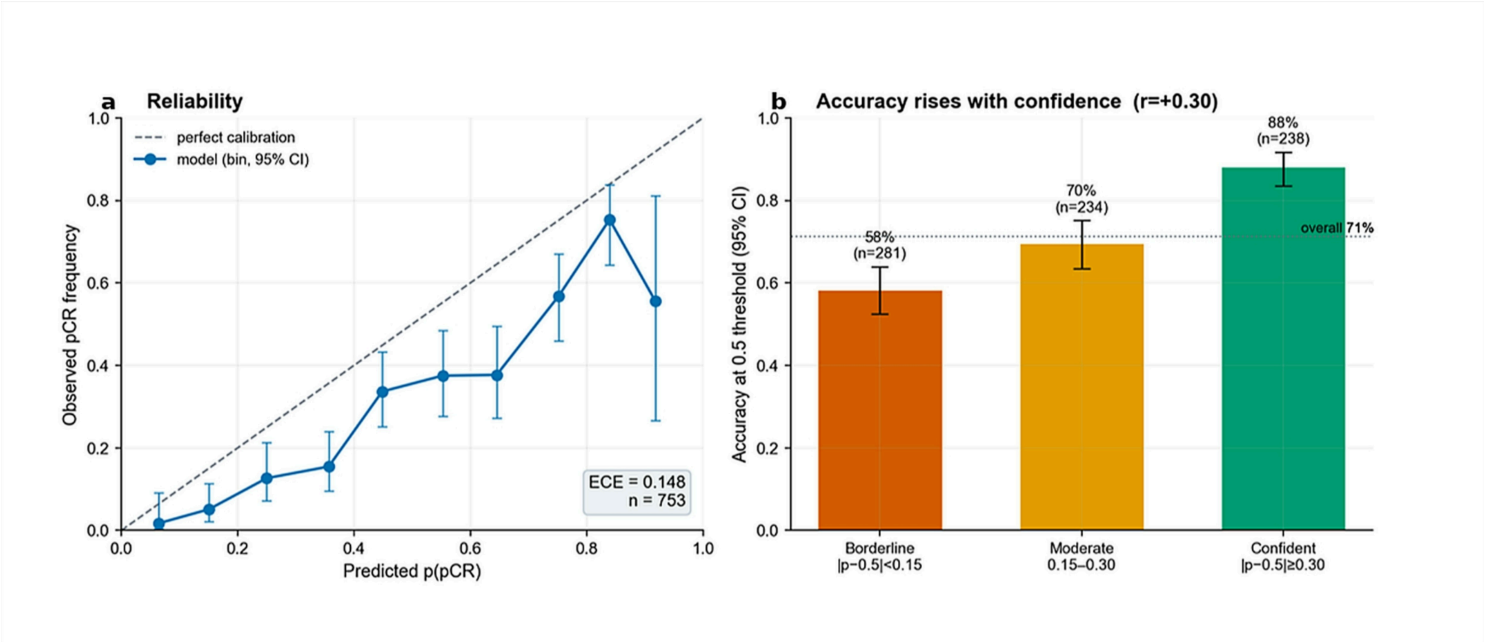


Figure 5

Model accuracy rises with prediction confidence. (a) Reliability diagram for the language-model readout with per-bin Wilson 95% confidence intervals (expected calibration error 0.148), showing overconfidence at intermediate predicted probabilities. (b) Accuracy was stratified by prediction confidence across the three confidence categories (borderline, moderate, and confident, with thresholds and sample sizes shown on the panel). Accuracy increased with confidence, and the confidence-accuracy correlation was $r = +0.30$. $n = 753$.

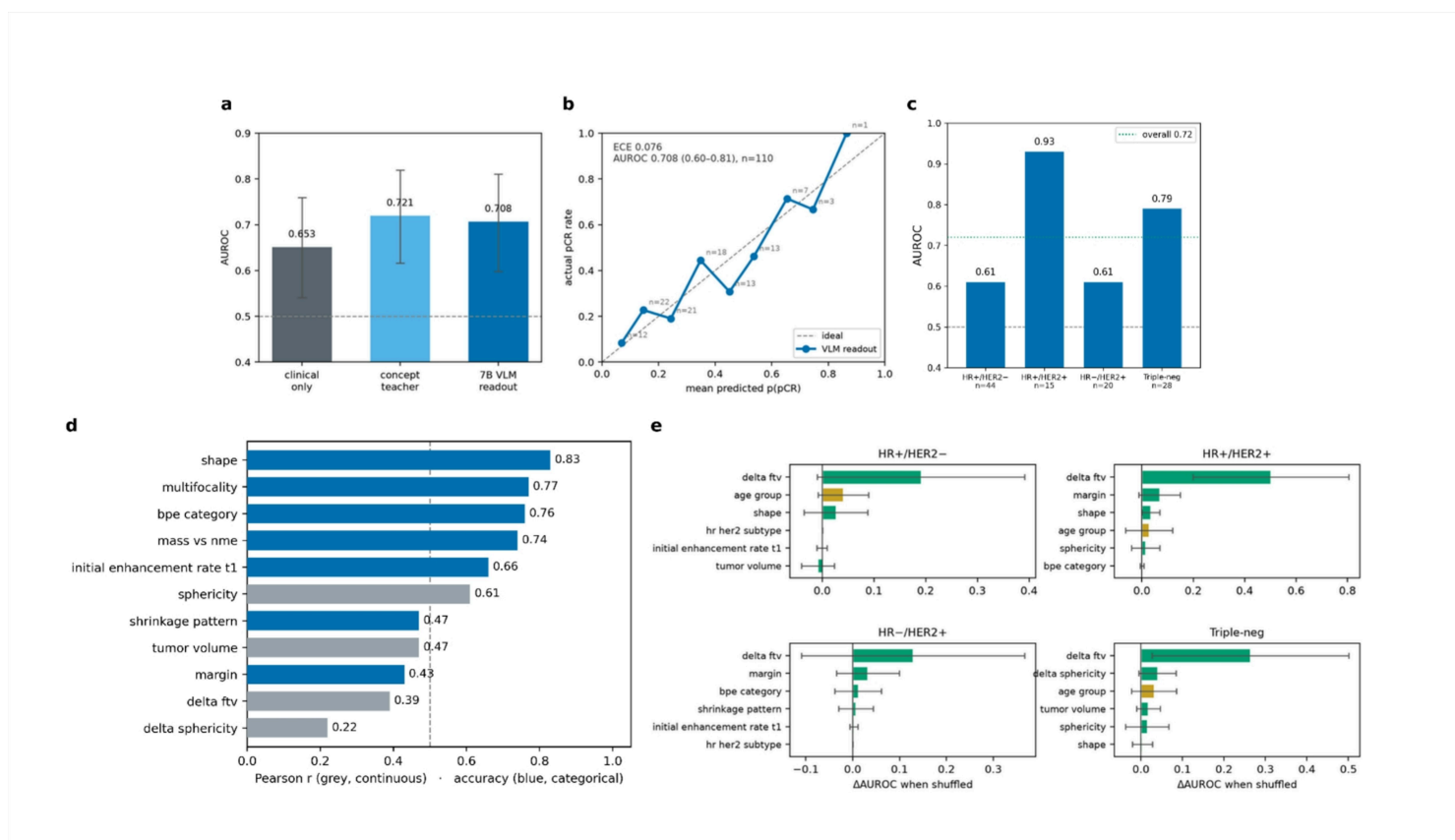


Figure 6

The concept-grounded response signal transferred to the independent I-SPY1/ACRIN-6657 cohort with preserved calibration and a stable concept ranking. The I-SPY2-trained model was applied unchanged to the external I-SPY1/ACRIN-6657 cohort. Because I-SPY1 delivered a single standard regimen, stratification by treatment mechanism was not applicable; treatment arm was not provided to the model in either cohort. (a) External discrimination was compared for the clinical-only model, the concept-and-clinical logistic teacher, and the 7B vision-language readout, with the corresponding internal I-SPY2 values shown for reference; the language readout did not retain its internal advantage over the teacher. (b) Reliability diagram for the language readout, with per-bin Wilson 95% confidence intervals (expected calibration error 0.077); calibration was preserved and numerically better than internally. (c) External AUROC by molecular subtype, with the overall value marked by the dotted line and per-subtype n shown on the panel; estimates in the smallest strata were unstable and are shown for completeness. (d) External concept fidelity for all eleven concepts, comparing predicted values against algorithmically derived reference values (Pearson r for continuous concepts, accuracy for categorical concepts). Fidelity was highest for the morphology concepts and lowest for the longitudinal-change concepts; the concepts that degraded most were those on which the internal model relied most heavily. (e) Per-variable permutation importance was computed within each molecular subtype (change in AUROC when a variable was shuffled; error bars ± 1.96 SD over permutation repeats). Change in functional tumor volume was the dominant early-response variable in every subtype, replicating the internal ranking. 115 quality-controlled subjects with paired T0 + T1 imaging were used for concept fidelity, which needs no outcome.

The 110 with an available pathologic complete response entered the 7B language-readout evaluation, and the 107 with complete pathologic complete response, molecular subtype, and age entered the concept-and-clinical and subtype-stratified analyses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.docx](#)