THERAROUTE-L: PATIENT-CONDITIONED TREATMENT-UNIT LEARNING FOR MULTIMODAL RESPONSE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Coarse treatment encodings leave drug composition and order implicit in multimodal prediction of pathological complete response (pCR) to neoadjuvant therapy. Learning patient-dependent interactions is further complicated by variable imaging histories and missing inputs. We introduce TheraRoute-L, jointly learning record-supported treatment-unit and relation representations under pCR supervision while retaining whole-regimen semantics. Clinical fields and observed MRI histories condition unit reads and bilinear scores of concurrent or sequential relations before fusion. A causal history encoder updates patient state, and seven private modality-subset experts combine available predictions. We derive interaction and read-sensitivity bounds and pooled and direct-read gradient paths, showing how direct reads bypass the zero-initialized residual map to supervise shared units. Retrospective evaluation spans four cohorts in two countries and multiple regions. With shared patient splits, labels, MRI preprocessing and scoring, baseline AUCs of 71.12% for I-SPY2 + Duke, 83.68% for Hospital-A and 79.56% for Hospital-B exceed those of retrained MRP–MRI and MORM in all three populations. Hospital-B AUC gains are 11.84% and 11.21%, respectively. Frozen expert restrictions support clinical–treatment complementarity. Code: `https://anonymous.4open.science/r/TheraRoute-L-6FBF/`.

1 INTRODUCTION

Multimodal treatment-response prediction faces a coupled representation and optimization problem: coarse regimen encodings leave drug composition and order implicit, while patient state, imaging history and input availability vary. Existing models combine clinical, imaging and treatment data (Han et al., 2025; Gao et al., 2024), but finer input granularity alone does not specify patient-dependent constituent interactions or how patient-level outcomes supervise shared treatment representations.

MRP integrates longitudinal imaging with clinical and treatment attributes (Gao et al., 2024), while MORM predicts response under recorded regimens (Han et al., 2025). Their treatment interfaces encode scalar attributes or regimen categories rather than explicit patient-conditioned unit-pair relations. Related work learns longitudinal representations (Gao et al., 2025; Ma et al., 2025a), uses MRI experts (Luo et al., 2025), or studies modality adaptation and routing (Zhang et al., 2024; Wu et al., 2025; Li et al., 2025). We address three coupled challenges: compositional interaction learning, history-dependent treatment reads, and prediction across available modality subsets.

TheraRoute-L learns shared representations of the finest record-supported units and their relations, anchored by whole-regimen semantics. Clinical fields and spatial–spectral MRI features condition unit reads and low-rank bilinear relation scores. A causal history encoder uses observed visits only; seven private modality-subset experts receive direct pCR supervision and use availability-masked routing.

The study makes three contributions:

1. **Patient-conditioned interaction modeling.** Shared bilinear maps encode symmetric concurrent and directed sequential relations. A mixed-difference bound characterizes three-way dependence beyond the defined pairwise-additive score class.

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.

2. **Outcome-supervised unit learning.** We derive pooled and direct-read gradients, including at zero residual initialization. Relation endpoints receive patient- and partner-dependent updates. Read sensitivity to scores is bounded regardless of set size.
3. **Cross-setting comparison.** Four retrospective cohorts across two countries and multiple regions provide 1,514 scored endpoints. Baseline AUCs reach 71.12% (I-SPY2 + Duke), 83.68% (Hospital-A) and 79.56% (Hospital-B), exceeding those of retrained MRP–MRI and MORM under shared splits, labels, MRI preprocessing and scoring; Hospital-B gains are 11.84% and 11.21%, respectively.

2 THERAROUTE-L

2.1 PREDICTION TASK AND MODEL OVERVIEW

For patient i at examination t, write $\widehat{p}_{i,t} = f_\theta(H_{i,\le t}, C_i, T_i, A_i)$. Here $H_{i,\le t}$ is the available MRI history, C_i is clinical context, $T_i = (w_i, U_i, E_i)$ contains whole-regimen text, supported units and recorded relations, and A_i specifies available modalities. Shared parameters θ learn from pCR reference labels. The three streams, denoted I, C and T, are encoded separately (Figure 1).

The *observed prefix* ends at the current examination. Source training uses these prefixes; performance tables use the first examination with retrospective clinical and treatment snapshots. The final implementation is called L3 in historical comparisons.

2.2 TREATMENT UNITS, RELATIONS AND PATIENT-CONDITIONED READS

Constructing the treatment input. Units retain the finest record-supported granularity: named medications, drug classes, alternative sets or unresolved regimens. Rule-based parsing prioritizes recorded arm components, otherwise using recognized expressions while preserving granularity and mechanism wording. Outcome supervision learns unit representations.

Recognized sequences yield directed edges; explicit concurrence links eligible named or class-level units. Separators alone create no edges. A documented concurrent docetaxel–doxorubicin–cyclophosphamide list yields three units and three relations; unresolved EC-T remains one regimen-level unit. Appendix B.4 gives cross-cohort examples and treatment-feature projections.

Encoding units and relations. Frozen PubMedBERT (Gu et al., 2021) encodes the whole description and each unit text. Padding is masked before pooling. Trainable maps combine these unit semantics with granularity, event-status and date-known embeddings to form a 128-dimensional feature u_j. Two shared projections produce 16-dimensional relation tokens:

$$e_{jk}^{\text{concurrent}} = (W_L u_j) \odot (W_L u_k), \qquad e_{jk}^{\text{sequential}} = (W_L u_j) \odot (W_R u_k). \tag{1}$$

Here $\odot$ is elementwise multiplication. Concurrent relations use the same projection at both endpoints and preserve the token when their order is swapped. Sequential relations use separate projections, allowing their tokens to depend on order.

The pooled treatment feature combines the whole description with the structured units and relations:

$$z^T = \phi_w(w) + \mathcal{R}(\overline{u} + W_e \overline{e}). \tag{2}$$

The bars denote valid-item means. ϕ_w projects the description, W_e projects relations, and $\mathcal{R}$ is a bias-free residual map initialized to zero. The pooled feature enters expert classification and routing. Direct patient reads below bypass $\mathcal{R}$. Appendix B.3 defines unit-only and empty-set cases.

Reading treatment in patient context. The joint treatment experts combine T with I (IT), C (CT), or both (ICT). Each forms a 16-dimensional patient query ξ_{ie}: image features supply IT, clinical features supply CT, and their concatenation supplies ICT. The query scores individual units and relations before pooling. For valid unit set J_i,

$$\begin{aligned} s_{ie,j}^U &= \xi_{ie}^\top W_{U,e} u_j / \sqrt{r}, \qquad \eta_{ie,j}^U = \frac{\exp(s_{ie,j}^U)}{\sum_{k \in J_i} \exp(s_{ie,k}^U)}, \\ v_{ie}^U &= \sum_{j \in J_i} \eta_{ie,j}^U \tanh(s_{ie,j}^U). \end{aligned} \tag{3}$$

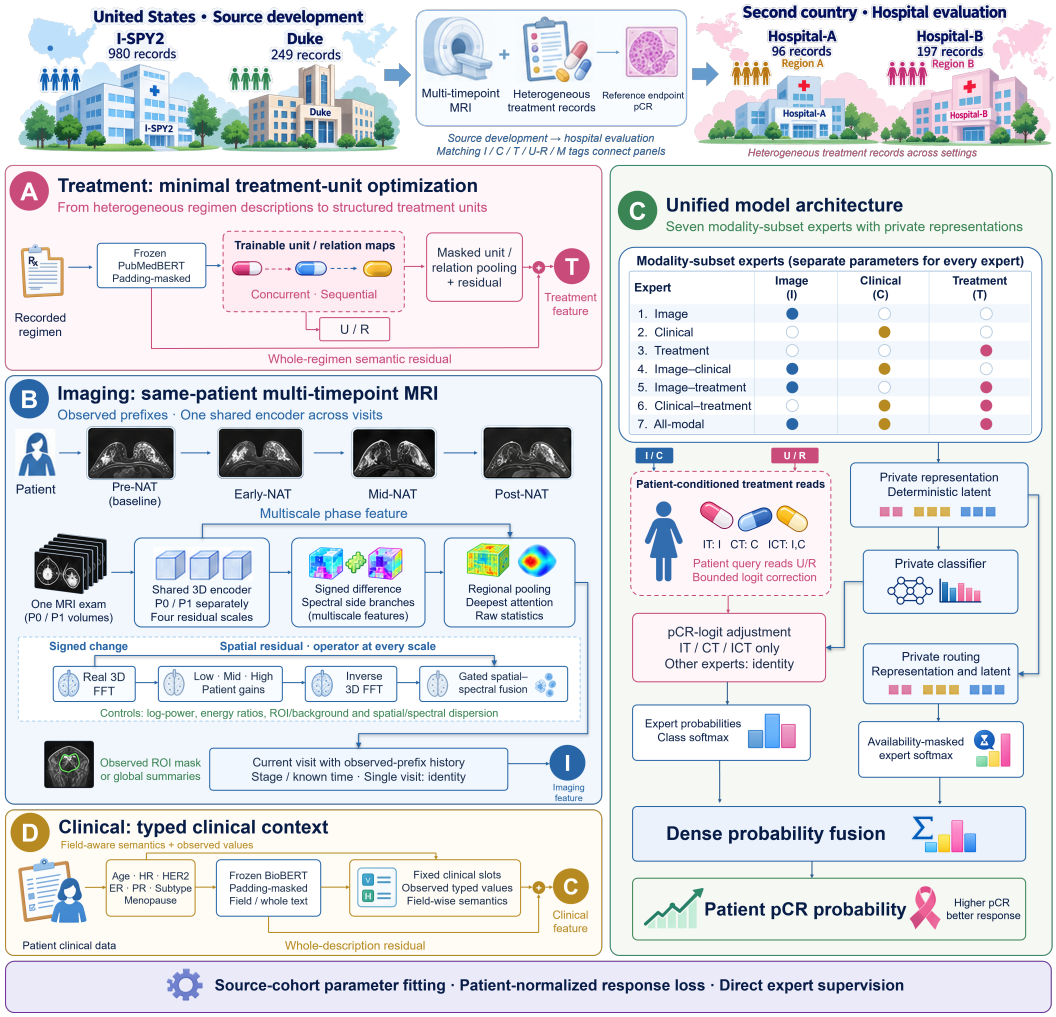


Figure 1: **TheraRoute-L architecture.** (A) Whole-regimen semantics combine with pooled, learned representations of treatment units and concurrent/sequential relations.(B) Shared P0/P1 spatial–spectral blocks, regional pooling, deepest-stage attention and raw statistics encode examinations; the history encoder uses available visits.(C) Seven private modality-subset experts have separate representations, classifiers and routing heads. IT/CT/ICT query units and relations using permitted patient modalities, yielding bounded pCR-logit corrections. Routing combines eligible expert probabilities.(D) Typed clinical fields combine observed values with frozen semantics. Top counts denote input records; visit icons are schematic. Performance tables use baseline MRI. Section 3.2 details source fitting and hospital evaluation.

The learned map $W_{U,e}$ projects the unit into the query space. Softmax normalizes the scores over valid units; their weighted $\tanh$ values give the scalar read v^U_{ie}. The relation read v^E_{ie} follows the same rule. Empty sets return zero.

The two reads produce a learned correction to the expert's positive-class logit, bounded in absolute value by $1/4$. Appendix B.3 gives its parameterization.

Shared structure, patient-specific scores. A sequential relation score can be written as

$$\begin{aligned} s^E_{ie,jk} &= u_j^\top B_{ie} u_k, \\ B_{ie} &= r^{-1/2} W_L^\top \operatorname{diag}(\xi_{ie}) W_R, \qquad \operatorname{rank}(B_{ie}) \le r, \quad r = 16. \end{aligned} \tag{4}$$

Writing the rows of the endpoint maps as $\ell_a^\top, h_a^\top$ gives $B(\xi) = r^{-1/2} \sum_a \xi_a \ell_a h_a^\top$. All patients therefore share a relation span of dimension at most r; their queries set its coefficients. Concurrent scores are symmetric, while sequential scores can depend on direction.

Proposition 1 (Patient–unit–unit interaction)**.** *For fixed endpoint maps, let $s(\xi, u, v) = u^\top B(\xi) v$. Choose two values for each latent input and let $\mathcal{Q}$ contain their eight combinations. With Δ denoting the difference between the two values,*

$$\begin{aligned} D := \Delta_\xi \Delta_u \Delta_v s = r^{-1/2} (\Delta\xi)^\top [(W_L \Delta u) \odot (W_R \Delta v)], \\ \max_{\mathcal{Q}} |s - g| \geq |D|/8 \quad \textit{for } g = a(\xi, u) + b(\xi, v) + c(u, v). \end{aligned} \tag{5}$$

Proof. Expanding the score gives $s = r^{-1/2} \sum_a \xi_a (\ell_a^\top u)(h_a^\top v)$. Each difference replaces its corresponding linear factor, yielding D. Every term of g omits one input, so its three-way difference is zero. The eight signed errors therefore satisfy

$$|D| = |\Delta_\xi \Delta_u \Delta_v (s - g)| \leq \sum_{q \in \mathcal{Q}} |s(q) - g(q)| \leq 8 \max_{q \in \mathcal{Q}} |s(q) - g(q)|. \tag{6}$$

Division by eight proves the claim. □

Thus $D \neq 0$ identifies a change in pair interaction with patient state. The score-level comparison concerns the stated additive class; A1/A2 retain nonlinear text and expert networks. The shared endpoint maps use 4,096 weights for $d = 128, r = 16$, independently of patient and regimen counts. Appendix A.1.1 details the latent-domain conditions.

Read stability and its derivation. Fix a nonempty valid set and finite scores, with numerical guards inactive. Write $\eta = \mathrm{softmax}(s)$ and $v = \sum_j \eta_j \tanh(s_j)$. Differentiating the weights and bounded values gives

$$\frac{\partial v}{\partial s_j} = \eta_j \{1 - \tanh^2(s_j) + \tanh(s_j) - v\}, \qquad \|\nabla_s v\|_1 \leq 3. \tag{7}$$

The bound follows from $1 - \tanh^2(s_j) \leq 1$, $|\tanh(s_j) - v| \leq 2$ and $\sum_j \eta_j = 1$. Integrating along the line from s to s' gives $|v(s) - v(s')| \leq 3\|s - s'\|_\infty$, independently of set size. For the implemented correction $r_e = \frac{1}{4} \tanh(a_e) \tanh(v^U + v^E)$, fixed masks and scale give $|r_e - r'_e| \leq \frac{3}{4}(\|s^U - s^{U\prime}\|_\infty + \|s^E - s^{E\prime}\|_\infty)$. Empty reads are zero; the formal inputs have zero interval reads (Appendix A.1.2).

2.3 MRI, EXAMINATION HISTORY AND CLINICAL CONTEXT

Within an MRI examination. P0 and P1 are contrast phases from one examination. A shared 3D residual encoder processes them separately. At each of four scales, a real Fourier branch modulates their signed P1–P0 feature difference over three radial frequency bands. A spatial residual combines with the modulated features. Regional descriptors, raw phase statistics and deepest-stage attention form the examination vector x_t. Tumor masks support ROI/background summaries; absent masks use global summaries. This branch applies spectral visual modeling (Lin et al., 2025) to signed within-examination changes. Appendix B specifies its dimensions, controller inputs and spectral bounds.

Across examinations. A history transformer uses stage and known relative-time inputs, attending only to current and earlier examinations. With its prefix state H_t and channel gate g_t, the image representation is

$$z_0 = x_0, \qquad z_t = x_t + g_t \odot \psi([x_0, x_t, x_t - x_0, x_t - x_{t-1}, H_t]), \quad t > 0. \tag{8}$$

The current examination has an identity path; baseline and adjacent-visit differences supply change information. Every argument is inside the observed prefix. Causal attention therefore makes earlier states independent of future visits when preprocessing and metadata stay fixed. Clinical and treatment snapshots are reused across stages.

For fixed treatment units and model parameters, two prefixes update the IT/ICT relation score through the patient query:

$$s^E_{t,jk} - s^E_{s,jk} = r^{-1/2} (\xi_t - \xi_s)^\top [(W_L u_j) \odot (W_R u_k)]. \tag{9}$$

Thus the same treatment structure is read against the updated imaging state. Concurrent relations replace W_R by W_L. Appendix A.6 derives query-sensitivity bounds and prefix consistency: future visits do not change an earlier prediction when its inputs and metadata are held fixed.

Clinical conditioning before fusion. Frozen BioBERT (Lee et al., 2020) semantics combine with observed values and identities in 32 fixed slots: age, HR, HER2, subtype, menopause and available ER/PR. Masked slots are concatenated and projected, with a whole-description residual of weight 0.1. In CT/ICT, clinical state enters the relation score through the query Jacobian:

$$\nabla_{z^C} s^E_{jk} = r^{-1/2} J^\top_{\xi, z^C}[(W_L u_j) \odot (W_R u_k)]. \tag{10}$$

This is the chain rule through the patient query, holding treatment and the other patient stream fixed. IT has no clinical input. Both text backbones stay in evaluation mode and mask padding before pooling; field dropout disables the whole-description bypass. Appendix B gives the slot construction.

2.4 Private experts and probability fusion

The seven experts use I, C, T, IC, IT, CT and ICT, respectively. Each private expert receives its declared modalities and has its own representation network, classifier and routing head. Experts using the same modality share its encoder. Treatment endpoint maps are also shared. IT, CT and ICT additionally use the patient-conditioned reads defined above.

Each eligible expert outputs a pCR probability p_{ie} and a routing score q_{ie}. The router takes a weighted mean of these probabilities:

$$\widehat{p}_i = \sum_{e \in \mathcal{S}(A_i)} \alpha_{ie} p_{ie}, \qquad \alpha_{ie} = \frac{\exp(q_{ie})}{\sum_{e' \in \mathcal{S}(A_i)} \exp(q_{ie'})}. \tag{11}$$

The set $\mathcal{S}(A_i)$ contains experts with all required inputs. All eligible experts participate; others have zero weight. At deterministic inference, removing a modality preserves the remaining experts' raw outputs. Their weights renormalize, with $|p - p_{\text{retained}}| \le \beta$ when the removed heads originally carry mass β and retained inputs stay fixed (Appendix D). This defines one prediction rule across observed availability patterns.

2.5 Training and patient-level weighting

Training starts with three epochs of image and clinical prediction. The next five epochs train all experts with uniform weights over available experts. Learned routing and checkpoint selection begin in epoch nine. The mixture objective is combined with direct expert supervision and modality dropout. The router also learns from a detached target based on the available experts' training cross-entropies. Related work distinguishes sufficient modality learning from equal modality influence (Ma et al., 2025b; Wei et al., 2025).

Weighting examination histories. A patient's baseline receives half of the supervised weight when valid follow-ups are present. The follow-ups share the other half. A baseline-only patient receives full weight. Appendix B defines the patient-level reduction and auxiliary-head eligibility.

Source weights balance the four cohort–outcome strata: the two US source cohorts crossed with pCR and non-pCR. Each stratum has equal total weight in the full-source primary prediction risk when every training patient has a valid mixture prediction. Appendix B.6.2 derives this identity and distinguishes the auxiliary-loss reductions.

From the classification objective to unit updates. Fix one patient-prefix, eligible experts, effective loss coefficients and stochastic realization. With numerical floors inactive, let $\pi_e = p_e^y (1 - p_e)^{1-y}$ be expert e's probability of the observed class. Its local classification objective and likelihood responsibility are

$$\mathcal{L}_{\text{pred}} = -\log \pi - \sum_e \lambda_e \log \pi_e, \qquad \pi = \sum_e \alpha_e \pi_e, \qquad \gamma_e = \frac{\alpha_e \pi_e}{\pi}. \tag{12}$$

For adjusted binary log-odds ζ_e, $\partial \log \pi_e / \partial \zeta_e = y - p_e$. Thus $\partial \mathcal{L}_{\text{pred}} / \partial \zeta_e = (\gamma_e + \lambda_e)(p_e - y)$. Direct supervision contributes independently of mixture responsibility. Differentiating learned softmax routing gives $\partial \mathcal{L}_{\text{pred}} / \partial q_e = \alpha_e - \gamma_e$.

Write ϑ for shared unit and endpoint parameters, $d_\vartheta = \overline{u} + W_e\overline{e}$, and R for the matrix of $\mathcal{R}$. For a record with whole text, the chain rule yields

$$\nabla_\vartheta \mathcal{L}_{\text{pred}} = \underbrace{J_{d_\vartheta}^\top R^\top g_T}_{\text{pooled feature}} + \underbrace{\sum_{e\in\mathcal{J}} c_e \nabla_\vartheta r_e}_{\text{patient-conditioned reads}}, \qquad c_e = (\gamma_e + \lambda_e)(p_e - y). \tag{13}$$

Here $g_T = \partial\mathcal{L}_{\text{pred}}/\partial z^T$ includes base-classifier and raw-router contributions; $\mathcal{J}$ contains eligible IT/CT/ICT experts. To derive the first term, differentiate $z^T = \phi_w(w) + Rd_\vartheta$ at fixed R. For the second, r_e enters only the positive logit, so its loss derivative is c_e. The patient query uses non-treatment modalities and is independent of ϑ.

At the implemented initialization $R = 0$, the pooled contribution to ϑ vanishes, but direct reads bypass R; $\nabla_R\mathcal{L}_{\text{pred}} = g_T d_\vartheta^\top$ updates the residual map. For a relation score, $\nabla_{u_j} s^E = B(\xi)u_k$ and $\nabla_{u_k} s^E = B(\xi)^\top u_j$: each endpoint receives a partner- and patient-dependent signal. Appendix A.2.1 covers unit-only inputs and warm-up stages.

Coupled learning across patients and regimens. Shared parameters accumulate classifier, router and direct-read signals across patients and prefixes. Each unit participates in its own read and incident relations, which also enter the pooled feature. Retraining a restricted pathway therefore changes coupled optimization across the system.

The full objective also includes router KL, image-attribute supervision, load balancing, and response/FFT regularization. Appendix B specifies all coefficients and reductions.

3 Data and evaluation

3.1 Four cohorts and heterogeneous treatment records

I-SPY2 ($N = 980$) and Duke ($N = 249$) form the US development group; Hospital-A ($N = 96$) and Hospital-B ($N = 197$) are anonymized hospitals in two regions of a second country. Physician-reviewed data comprise 1,522 records and 1,514 evaluable endpoints. Eight unresolved Hospital-B endpoints are excluded from scoring and threshold fitting. Appendices B.1 and G give cohort counts, input composition and endpoint sources.

Treatment records span named drugs, classes, alternatives and unresolved regimens, with cohort-dependent concurrent and sequential relations (Table 4). Development and Hospital-A have baseline masks and usable clinical and treatment branches. Hospital-B has no masks; 187 of its 189 scored records have clinical representations and 35 have treatment representations.

Follow-up MRI from 582 I-SPY2 patients enters source training; other cohorts contribute baseline inputs. All performance tables use baseline MRI and retrospective clinical/treatment snapshots, whose pretreatment availability was unverified. Treatment availability is outcome-associated.

3.2 Fitting, hospital use and decision thresholds

Five patient-level outer folds stratify source records by cohort and label (seed 42), each with 835/836 training, 148 inner-validation and 246/245 outer records. Inner validation selects checkpoints and cohort-by-task Youden cutoffs; development pools one outer-held-out score per record. Hospitals use equal-weight means of five frozen source models. Hospital-A uses cutoff 0.5; Hospital-B fits Youden cutoffs on four local stratified folds and applies them to the fifth (locked order, seed 42). Network parameters and ensemble weights remain fixed. Both hospitals informed system development; evaluation is retrospective.

AUC uses continuous probabilities; ACC, SEN, SPE, PPV and NPV use pooled held-out decisions, with undefined predictive values shown as N/A. All 95% intervals use 1,000 class-stratified bootstrap resamples conditional on fixed predictions and decisions. Paired AUC differences share resampling indices and have no multiplicity adjustment. Appendices B.7 and C detail selection and scoring.

3.3 Controlled ablations and comparator protocols

Published models. Retrained MRP–MRI (Gao et al., 2024) and MORM (Han et al., 2025) share source partitions, labels, MRI preprocessing and scoring rules, retaining their MRI–scalar and multimodal product-of-experts architectures (Appendix F.2).

Controlled retraining. A1 uses whole-regimen text; A2 adds units and their reads; A0 also adds relation pooling and reads. Five-fold runs fix patient inputs, shared-parameter initialization, training and checkpoint-selection rules. Hospital-B uses variant-specific cutoffs within common local folds (Appendix F.1).

Historical systems. L1/L2 and static V1/V2/V4/V5 retain their original training and evaluation procedures (Appendix F).

Frozen restrictions. Removing I-, C- or T-containing experts retains C/T/CT, I/T/IT or I/C/IC; single-modality restrictions retain one head. Parameters and missing inputs stay fixed; retained weights renormalize before ensembling. Restriction-specific cutoffs use the original validation partitions (Appendix D).

4 Results

4.1 Comparison with published multimodal response models

TheraRoute-L achieves the highest baseline AUC in all three populations compared with retrained MRP–MRI (Gao et al., 2024) and MORM (Han et al., 2025) (Table 1). Gains over the stronger comparator are 2.28 percentage points in development, 5.92 in Hospital-A and 11.21 in Hospital-B, calculated from unrounded estimates.

Hospital-B shows the largest separation: AUC is 79.56% versus 67.72% for MRP–MRI and 68.35% for MORM; accuracy gains are 9.52 and 14.81 percentage points. TheraRoute-L also leads in specificity and PPV there, while MORM has higher sensitivity and NPV. In Hospital-A, MORM has slightly higher accuracy despite lower AUC. MRP–MRI predicts no positives at 0.5, so its PPV is undefined.

4.2 Joint prediction across available modality subsets

Table 2 restricts the same frozen model, renormalizing retained-head weights while keeping encoders, classifiers and raw routing scores fixed. It describes information use within the jointly trained system, complementing the independently retrained model comparisons in Table 1.

Clinical context and treatment work together. The C/T/CT restriction exceeds C-only AUC on matched records in development, Hospital-A and Hospital-B, by 2.08, 2.94 and 3.16 percentage points, respectively. Conversely, removing clinical-containing experts gives the weakest of the three modality-removal configurations in every population. Removing treatment-containing experts also lowers AUC relative to the full mixture in all three populations. These patterns support combining patient context with treatment information inside the learned predictor.

Coverage and discrimination answer different questions. C/T/CT also exceeds the matched full-mixture AUCs of 71.12%, 83.68% and 80.07%. Thus the available baseline MRI does not improve every retained-subset comparison. In Hospital-B, however, the image expert supplies predictions for two additional records with neither usable clinical nor treatment representations, allowing the full model to cover all 189 endpoints. Its image branch also supplies the changing patient state when longitudinal MRI is available during source training.

Hospital-B T-only covers 35 records: its AUC 72.35% compares with 90.91% for the full mixture on those same patients. The 187-record C/T/CT comparison uses a different population. Treatment availability is outcome-associated, so eligibility must accompany these estimates. Appendix D gives the restriction algebra and complete metrics; Appendix E illustrates patient-conditioned reads.

Table 1: **Comparison with published multimodal response models under a shared evaluation protocol.** Parentheses: 95% conditional bootstrap intervals; classification metrics: percentages. Development uses outer-held-out scores and source-inner cutoffs. Hospitals use five-model means: A at 0.5, B with cross-fitted Youden cutoffs. Bold: column maxima. †Retrained adaptations: MRP's MRI branch (Gao et al., 2024) and MORM (Han et al., 2025); details in Appendix F.2. N/A: undefined PPV and interval with no predicted positives.

Model	AUC	ACC	SEN	SPE	PPV	NPV
I-SPY2 + Duke ($N = 1,229$)						
MRP–MRI†	68.84% (65.43%–71.97%)	62.90 (60.13–65.58)	54.55 (49.27–60.12)	**66.10** (63.06–69.37)	38.19 (35.02–41.42)	79.11 (77.06–81.22)
MORM†	64.51% (61.22%–67.97%)	61.35 (58.67–64.12)	52.49 (46.92–57.77)	64.75 (61.71–67.91)	36.38 (33.20–39.55)	78.02 (75.85–80.16)
TheraRoute-L	**71.12%** (67.93%–74.26%)	**65.58** (62.89–68.35)	**70.09** (65.40–74.78)	63.85 (60.59–67.12)	**42.68** (40.03–45.65)	**84.75** (82.57–86.92)
Hospital-A ($N = 96$)						
MRP–MRI†	73.80% (63.58%–83.59%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	**100.00** (100.00–100.00)	N/A	55.21 (55.21–55.21)
MORM†	77.75% (68.10%–86.84%)	**77.08** (68.75–84.38)	**83.72** (72.09–93.02)	71.70 (58.49–83.02)	70.59 (61.40–80.50)	**84.44** (75.56–93.76)
TheraRoute-L	**83.68%** (75.12%–91.18%)	76.04 (67.71–84.38)	79.07 (67.44–90.70)	73.58 (62.22–84.91)	**70.83** (61.11–81.40)	81.25 (72.92–90.91)
Hospital-B ($N = 189$)						
MRP–MRI†	67.72% (58.83%–75.97%)	67.20 (59.79–74.07)	69.74 (59.21–78.95)	65.49 (56.64–74.34)	57.61 (50.00–65.56)	76.29 (69.57–82.98)
MORM†	68.35% (60.39%–75.82%)	61.90 (55.03–68.25)	**84.21** (75.00–92.11)	46.90 (37.17–55.75)	51.61 (46.51–57.03)	**81.54** (72.97–90.16)
TheraRoute-L	**79.56%** (72.89%–85.78%)	**76.72** (70.37–82.54)	68.42 (57.89–77.63)	**82.30** (74.34–89.38)	**72.22** (63.75–80.89)	79.49 (74.19–84.96)

Table 2: **Joint prediction and coverage under frozen expert restrictions.** Listed heads are intersected with each record's available inputs; all rows use the same trained parameters. Parentheses: 95% conditional AUC intervals. Development uses outer-held-out scores; hospitals use five-model means after restriction. N_B gives Hospital-B eligibility. The final two rows report recorded matched-population point estimates. Full six-metric results are in Appendix D.

Restriction	Retained heads	I-SPY2 + Duke $N = 1,229$	Hospital-A $N = 96$	Hospital-B	N_B
Full	All seven	71.12% (67.93%–74.26%)	83.68% (75.12%–91.18%)	79.56% (72.89%–85.78%)	189
Without I	C/T/CT	71.77% (68.73%–74.82%)	85.26% (77.18%–92.19%)	80.92% (74.26%–86.91%)	187
Without C	I/T/IT	65.04% (61.85%–68.33%)	77.31% (66.60%–86.97%)	54.98% (46.03%–64.04%)	189
Without T	I/C/IC	69.61% (66.51%–72.80%)	79.42% (69.85%–88.24%)	76.58% (68.93%–83.14%)	189
I only	I	56.18% (52.86%–59.66%)	48.66% (36.99%–60.21%)	58.77% (50.41%–67.18%)	189
C only	C	69.69% (66.55%–72.75%)	82.32% (73.54%–89.82%)	77.76% (70.89%–84.20%)	187
T only	T	66.30% (63.15%–69.56%)	79.55% (69.68%–87.89%)	72.35% (54.92%–87.88%)	35
Full mixture on the C-eligible Hospital-B records				80.07%	187
Full mixture on the T-eligible Hospital-B records				90.91%	35

4.3 JOINTLY LEARNED TREATMENT REPRESENTATIONS

Within a common architecture, A1/A2/A0 yield small source AUC differences and population-dependent hospital increments. These findings complement the coupled gradient paths in Equation (13): regimen semantics, units and relations form jointly learned representations. They are consistent with complementary roles, without establishing large standalone module gains. Appendix F.1 retains all AUCs, six-metric results and paired intervals (Tables 11–15).

4.4 COMPARISON WITH HISTORICAL COMPLETE SYSTEMS

Among the seven historical systems, TheraRoute-L has the highest Hospital-B AUC, ACC, PPV and NPV and exceeds L2 in all six point estimates (Table 3). Its AUC also exceeds L1 and L2 in all three populations; gains over L2 are 1.43, 2.42 and 11.37 percentage points for development, Hospital-A and Hospital-B (Appendix F; Supplementary Figure S3a). Historical systems retain their original training configurations. These results evaluate complete designs with several interacting changes; the within-L3 treatment comparisons are reported separately in Appendix F.1.

Table 3: **Hospital-B comparison of complete systems under a common scoring protocol.** All systems use the same 189 endpoints, equal-weight five-model probabilities and local five-fold cross-fitted Youden cutoffs. AUC uses continuous probabilities; other metrics are percentages. Parentheses give 95% conditional bootstrap intervals; bold marks column maxima. Source training, preprocessing and calibration follow each original system (Appendix F).

System	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
Static V1	51.46% (42.82%–59.35%)	50.26 (43.92–56.61)	26.32 (15.79–36.84)	66.37 (57.52–74.34)	34.48 (23.73–44.64)	57.25 (52.67–61.79)
Static V2	62.23% (53.94%–69.70%)	59.79 (52.91–66.14)	39.47 (28.95–50.00)	73.45 (64.60–80.53)	50.00 (39.19–60.61)	64.34 (59.26–69.17)
Static V4	47.28% (39.19%–55.70%)	52.91 (46.03–59.79)	71.05 (61.84–81.58)	40.71 (31.86–49.56)	44.63 (39.67–50.00)	67.65 (58.73–76.92)
Static V5	59.05% (50.67%–67.30%)	63.49 (58.19–68.80)	32.89 (22.37–43.42)	**84.07** (77.88–90.27)	58.14 (46.00–70.73)	65.07 (61.43–69.18)
L1	54.02% (45.80%–62.86%)	52.38 (45.50–59.26)	**73.68** (63.16–84.21)	38.05 (29.20–46.92)	44.44 (39.52–49.60)	68.25 (58.46–78.95)
L2	68.19% (60.54%–76.10%)	62.43 (55.56–69.31)	55.26 (44.74–65.82)	67.26 (58.41–76.11)	53.16 (45.12–61.97)	69.09 (63.55–75.42)
TheraRoute-L (L3)	**79.56%** (72.89%–85.78%)	**76.72** (70.37–82.54)	68.42 (57.89–77.63)	82.30 (74.34–89.38)	**72.22** (63.75–80.89)	**79.49** (74.19–84.96)

5 DISCUSSION AND CONCLUSION

TheraRoute-L combines heterogeneous treatment records and patient observations through jointly learned representations. Its score and gradient analyses explain how patient state, units and relations participate in one predictor; the empirical evidence evaluates their combined performance. The modest treatment-ablation differences are consistent with this integrated design.

Frozen restrictions distinguish subset discrimination from record coverage. Clinical–treatment predictions are strong across the three populations, while the full model covers records with missing clinical/treatment inputs. Longitudinal value should be evaluated on matched patients across stages.

These findings concern retrospective prediction under documented regimens. Pretreatment availability of contextual snapshots was unverified, treatment availability was outcome-associated, and both hospitals informed system development. Prospective evaluation with time-verified inputs, broader treatment coverage and matched baseline/follow-up populations is needed to establish clinical utility, assess longitudinal prediction across institutions and examine how performance varies with treatment granularity and documentation practices across clinical settings.

AI USE STATEMENT

AI assistance was used for implementation and result inspection, manuscript drafting, editing and formatting, data summarization and aggregate plotting, numerical checks, verification of mathematical derivations, and literature searches. Reported performance metrics were calculated from stored model predictions and study reference endpoints. The authors take responsibility for final verification and comprehensive disclosure of AI use throughout the project. No local patient data were transmitted to public literature-search services.

ETHICS STATEMENT

This retrospective study uses physician-reviewed data and reference annotations. The Hospital-A and Hospital-B datasets were collected and annotated by physicians at the respective institutions. Hospital identities are anonymized, and Hospital-B records with unresolved endpoints were excluded from scoring and threshold fitting. Raw archives, identifiable records and patient-level prediction tables remain under restricted access. Manuscript illustrations use de-identified derived image slices. Institutional approvals were obtained, and the corresponding approval identifiers and informed-consent or waiver documentation were finalized before submission.

REPRODUCIBILITY STATEMENT

The model equations and architecture are specified in Section 2, with structural and optimization proofs provided in Appendix A. Section 3 and Appendices B–D define input construction, training, folds, thresholds, uncertainty estimation and frozen-expert restrictions. Appendix F reports controlled ablations, the retrained published-model adaptations and historical configurations. Supplementary treatment-feature projections use a fixed source checkpoint, de-duplicated regimen inputs and recorded projection settings. The anonymous supplementary archive contains the implementation and scoring guide, code/configuration fingerprints, frozen language-model and tokenizer specifications, and numerical and automatic-differentiation checks for the derivations. Metrics are traceable to fixed analysis populations, frozen predictions and saved decisions. Appendix G documents endpoint provenance and physician confirmation of the 28 Duke reference labels. Figures use aggregate results or de-identified derived images only. Source code is available at `https://anonymous.4open.science/r/TheraRoute-L-6FBF/`. Hospital-A and Hospital-B datasets are not publicly released; patient-level data remain under institutional access controls.

REFERENCES

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24507–24517, 2025. URL `https://www.openaccess.thecvf.com/content/ICCV2025/html/Dai_Unbiased_Missing-modality_Multimodal_Learning_ICCV_2025_paper.html`.

Yaofei Duan, Yuhao Huang, Xin Yang, Luyi Han, Xinyu Xie, Zhiyuan Zhu, Ping He, Ka-Hou Chan, Ligang Cui, Sio-Kei Im, Dong Ni, and Tao Tan. ADAptation: Reconstruction-based unsupervised active learning for breast ultrasound diagnosis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15975 of *Lecture Notes in Computer Science*, pp. 35–45. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05325-1_4. URL `https://doi.org/10.1007/978-3-032-05325-1_4`.

Yuan Gao, Sofia Ventura-Diaz, Xin Wang, Muzhen He, Zeyan Xu, Arlene Weir, et al. An explainable longitudinal multi-modal fusion model for predicting neoadjuvant therapy response in women with breast cancer. *Nature Communications*, 15:9613, 2024. doi: 10.1038/s41467-024-53450-8. URL `https://doi.org/10.1038/s41467-024-53450-8`.

Yuan Gao, Tao Tan, Xin Wang, et al. Multi-modal longitudinal representation learning for predicting neoadjuvant therapy response in breast cancer treatment. *IEEE Journal of Biomedical and*

Health Informatics, 29(12):9041–9050, 2025. doi: 10.1109/JBHI.2025.3540574. URL https://doi.org/10.1109/JBHI.2025.3540574.

Yuan Gao, Hong-Yu Zhou, Xin Wang, Antonio Portaluri, Tianyu Zhang, Regina Beets-Tan, Luyi Han, Chunyao Lu, Laura Estacio, Anna D'Angelo, Stephan Ursprung, Y. Yu, Jonas Teuwen, Tao Tan, and Ritse Mann. Visualizing radiologic connections: An explainable coarse-to-fine foundation model with multiview mammograms and associated reports. *Radiology: Artificial Intelligence*, 8(1):e240646, 2026. doi: 10.1148/ryai.240646. URL https://doi.org/10.1148/ryai.240646.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Luyi Han, Tianyu Zhang, Anna D'Angelo, et al. Exploring personalized neoadjuvant therapy selection strategies in breast cancer: an explainable multi-modal response model. *eClinicalMedicine*, 86: 103356, 2025. doi: 10.1016/j.eclinm.2025.103356. URL https://doi.org/10.1016/j.eclinm.2025.103356.

Luyi Han, Tao Tan, Tianyu Zhang, Yuan Gao, Xin Wang, Valentina Longo, Sofía Ventura-Díaz, Anna D'Angelo, Yue Sun, Jonas Teuwen, and Ritse Mann. Tumor-preserving deformable registration of longitudinal DCE breast MRI during neoadjuvant chemotherapy. *Radiology: Artificial Intelligence*, 8(4):e250789, 2026. doi: 10.1148/ryai.250789. URL https://doi.org/10.1148/ryai.250789.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Zhongyang Li, Ziyue Li, and Tianyi Zhou. R2-T2: Re-routing in test-time for multimodal mixture-of-experts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35292–35316. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25bc.html.

Xinglong Liang, Tianyu Zhang, Miguel Braga, Luyi Han, Maarten Donswijk, Jiajun Huang, Jinhong Song, Chunyao Lu, Xin Wang, Yuan Gao, Chunyan Xiong, Yue Sun, Jun Xu, Jonas Teuwen, Wouter Vogel, Tao Tan, and Ritse Mann. Multi-omics deep learning improves FDG PET-CT-based long-term prognostication of breast cancer. *npj Precision Oncology*, 10:74, 2026. doi: 10.1038/s41698-026-01283-7. URL https://doi.org/10.1038/s41698-026-01283-7.

Wang Lin, QingSong Wang, Yueying Feng, Shulei Wang, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, and Jingyuan Chen. Non-natural image understanding with advancing frequency-based vision encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29756–29766, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lin_Non-Natural_Image_Understanding_with_Advancing_Frequency-based_Vision_Encoders_CVPR_2025_paper.html.

Chunyao Lu, Tianyu Zhang, et al. LUMIN: A longitudinal multi-modal knowledge decomposition network for predicting breast cancer recurrence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7530–7538, 2026. doi: 10.1609/aaai.v40i9.37693. URL https://doi.org/10.1609/aaai.v40i9.37693.

Luyang Luo, Mingxiang Wu, Mei Li, Yi Xin, Qiong Wang, Varut Vardhanabhuti, Winnie CW Chu, Zhenhui Li, Juan Zhou, Pranav Rajpurkar, and Hao Chen. A large model for non-invasive and personalized management of breast cancer from multiparametric MRI. *Nature Communications*, 16:3647, 2025. doi: 10.1038/s41467-025-58798-z. URL https://www.nature.com/articles/s41467-025-58798-z.

Dingrui Ma, Jiawei Cao, Hao Cheng, Dan Zhou, Jianping Liu, Xiaofeng Zhang, Kaijie Wu, Chaochen Gu, and Xinping Guan. Longitudinal MRI-clinical multimodal fusion for pCR prediction in

breast cancer. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15974 of *Lecture Notes in Computer Science*, pp. 323–332. Springer, 2025a. doi: 10.1007/978-3-032-05182-0_32. URL https://papers.miccai.org/miccai-2025/0510-Paper2501.html.

Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41827–41836. PMLR, 2025b. URL https://proceedings.mlr.press/v267/ma25c.html.

Jason Qin, Hans-Hermann Wessels, Carlos Fernandez-Granda, and Yuhan Hao. Active learning for efficient discovery of optimal combinatorial perturbations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50363–50383. PMLR, 2025. URL https://proceedings.mlr.press/v267/qin25g.html.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Xin Wang, Tao Tan, Yuan Gao, Ruisheng Su, Jonas Teuwen, Jaap Kroes, Tianyu Zhang, Anna D'Angelo, Luyi Han, Caroline A. Drukker, Marjanka K. Schmidt, Regina Beets-Tan, Nico Karssemeijer, and Ritse Mann. Predicting short- to long-term breast cancer risk from longitudinal mammographic screening history. *npj Breast Cancer*, 11:118, 2025a. doi: 10.1038/s41523-025-00831-x. URL https://doi.org/10.1038/s41523-025-00831-x.

Xin Wang, Tao Tan, Yuan Gao, Hong-Yu Zhou, Tianyu Zhang, Luyi Han, Antonio Portaluri, Eric Marcus, Chunyao Lu, Caroline A. Drukker, Jonas Teuwen, Regina Beets-Tan, Shanshan Wang, Nico Karssemeijer, and Ritse Mann. Mammo-AGE: deep learning estimation of breast age from mammograms. *Nature Communications*, 16:10934, 2025b. doi: 10.1038/s41467-025-65923-5. URL https://doi.org/10.1038/s41467-025-65923-5.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2250–2259, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wei_Improving_Multimodal_Learning_via_Imbalanced_Learning_ICCV_2025_paper.html.

Chenwei Wu, Zitao Shuai, Zhengxu Tang, Luning Wang, and Liyue Shen. Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e3b82f4c7ba93a88025cf97dca9edc83-Abstract-Conference.html.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Michael Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f49d76cf84df83a611883c621c96d2d9-Abstract-Conference.html.

Tianyu Xie, Yue Sun, Hongxu Yang, et al. Semisupervised breast MRI density segmentation integrating fine and rough annotations. *IEEE Transactions on Artificial Intelligence*, 6(3):690–699, 2025. doi: 10.1109/TAI.2024.3491693. URL https://doi.org/10.1109/TAI.2024.3491693.

Xinyu Xie, Luyi Han, Yonghao Li, et al. Tumor segmentation with heterogeneity clustering in non-contrast breast MRI. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15961 of *Lecture Notes in Computer Science*, pp. 642–652. Springer, 2026. doi: 10.1007/978-3-032-04937-7_61. URL https://doi.org/10.1007/978-3-032-04937-7_61.

Siyi Xun, Yue Sun, Jingkun Chen, Zitong Yu, Tong Tong, Xiaohong Liu, Mingxiang Wu, and Tao Tan. MedIQA: A scalable foundation model for prompt-driven medical image quality assessment. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15972 of *Lecture Notes in Computer Science*, pp. 339–349. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05169-1_33. URL https://doi.org/10.1007/978-3-032-05169-1_33.

Hao Yang, Xinglong Liang, Zhang Li, Yue Sun, Zheyu Hu, Ritse Mann, Qifeng Yu, and Tao Tan. Prompt mechanisms in medical imaging: A comprehensive survey. *The Innovation*, 7(6): 101271, 2026. doi: 10.1016/j.xinn.2026.101271. URL https://doi.org/10.1016/j.xinn.2026.101271.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27456–27466, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Multimodal_Representation_Learning_by_Alternating_Unimodal_Adaptation_CVPR_2024_paper.html.

Zhuoneng Zhang, Luyi Han, Dengqiang Jia, et al. MRICombo: a deep-learning-based framework for universal volumetric segmentation grading-staging and malignancy detection across heterogeneous MRI. *Nature Communications*, 2026. doi: 10.1038/s41467-026-76461-z. URL https://doi.org/10.1038/s41467-026-76461-z.

Dashun Zheng, Patrick Cheong-Iao Pang, Ping He, et al. Parallel multimodal language model: Enhanced breast nodule diagnosis through parallel multimodal representations and large language models. *Advanced Intelligent Systems*, 8(1):2500085, 2026. doi: 10.1002/aisy.202500085. URL https://doi.org/10.1002/aisy.202500085.

A Structural and optimization properties

We analyze the relation parameterization, computational dependencies and supervised optimization of TheraRoute-L. Each result specifies its fixed quantities and operating conditions. Derivatives are taken in real arithmetic with the relevant numerical guards inactive.

A.1 Shared relation components and conditional non-additivity

Proposition 2 (Contextual bilinear relation)**.** *Let unit features have dimension d and endpoint projections have dimension r. The sequential score matrix B_{ie} in Equation* (4) *then has rank at most r. Concurrent scores are symmetric in their endpoints. Independent left/right projections allow directional sequential scores.*

Let $\ell_a^\top$ and $h_a^\top$ be row a of W_L and W_R. The patient-conditioned matrix has the expansion

$$B_{ie} = r^{-1/2} \sum_{a=1}^{r} \xi_{ie,a} \ell_a h_a^\top. \tag{14}$$

Proposition 3 (Conditional interaction structure)**.** *For a fixed patient query ξ, let $s_\xi(u, v) = u^\top B(\xi) v$. For any latent endpoints,*

$$\begin{aligned} s_\xi(u, v) - s_\xi(u', v) - s_\xi(u, v') + s_\xi(u', v') \\ = (u - u')^\top B(\xi)(v - v'). \end{aligned} \tag{15}$$

If $B(\xi) \neq 0$, the score cannot be written as $a(u)+b(v)$ over the full latent space. Concurrent relations use W_L at both endpoints and are symmetric; independent sequential maps permit orientation differences $u^\top(B - B^\top)v$.

Proof of the factorization and Proposition 3. The native score is

$$r^{-1/2} \sum_{a=1}^{r} \xi_{ie,a} (W_L u_j)_a (W_R u_k)_a = u_j^\top \{ r^{-1/2} W_L^\top \operatorname{diag}(\xi_{ie}) W_R \} u_k.$$

Expanding the two endpoint matrices by their rows gives $r^{-1/2}\sum_a \xi_{ie,a}\ell_a h_a^\top$, as in Equation (14). Each shared matrix has rank at most one; these matrices need not be linearly independent. The patient-conditioned matrices therefore lie in a shared linear span of dimension at most r. Each also has rank at most r. The query's $\tanh$ coordinates bound the combination coefficients within this span.

For fixed query ξ, expanding the four bilinear terms gives Equation (15). Every separable additive function $a(u)+b(v)$ has zero mixed difference because its two constituent functions cancel separately. If $B(\xi)\neq 0$, choose an entry $B_{ab}\neq 0$ and coordinate differences $u-u'=\mathbf{e}_a$, $v-v'=\mathbf{e}_b$. The mixed difference is then $B_{ab}\neq 0$, proving that the score has no globally separable additive representation on the full latent endpoint space.

If $W_R = W_L$, transposition leaves the middle matrix unchanged, so the endpoint score is symmetric. Otherwise reversing the endpoints changes the score by $u_j^\top(B_{ie}-B_{ie}^\top)u_k$, which can be nonzero. Because ξ_{ie} is produced from the expert's permitted patient modalities, this middle matrix can vary across patients even for identical treatment units. The two projection matrices contain $2rd$ weights independent of the number of complete regimens or patients. □

For $d = 128$ and $r = 16$, the two bias-free endpoint maps contain 4,096 weights. The semantic encoder and expert-specific query networks contribute additional parameters. Signed patient-query coefficients give symmetric, potentially indefinite concurrent score matrices. Proposition 3 establishes conditional non-additivity of the learned latent relation scores. Documented endpoints and types instantiate this structure. Source inputs contain sequential relations and zero explicit concurrent relations. The shared left map learns from sequential edges, and its symmetric extension processes concurrent hospital edges.

A.1.1 Patient–unit–unit interaction and approximation gap

Proposition 1 analyzes the relation score as a function of its latent patient query and two latent treatment endpoints. It concerns the explicit score operator, with learned endpoint maps fixed.

Proof of Proposition 1. Let ξ^0,ξ^1, u^0,u^1 and v^0,v^1 define the eight combinations in $\mathcal{Q}$. For any function h, define

$$\Delta_\xi\Delta_u\Delta_v h = \sum_{a,b,c\in\{0,1\}} (-1)^{3-a-b-c} h(\xi^a,u^b,v^c). \tag{16}$$

Use $\Delta\xi = \xi^1-\xi^0$ and analogously for the endpoints. The score has the expansion

$$s(\xi,u,v) = r^{-1/2}\sum_{j=1}^{r}\xi_j(\ell_j^\top u)(h_j^\top v). \tag{17}$$

Taking each difference in turn replaces the corresponding linear factor by its difference. This yields the expression for D in Equation (5). Every term of $g = a(\xi,u)+b(\xi,v)+c(u,v)$ is independent of one input, so the difference in that input cancels the term. Hence $\Delta_\xi\Delta_u\Delta_v g = 0$, and

$$|D| = |\Delta_\xi\Delta_u\Delta_v(s-g)| \le \sum_{a,b,c}|(s-g)(\xi^a,u^b,v^c)| \le 8\max_{\mathcal{Q}}|s-g|.$$

Rearranging proves the gap. Concurrent relations use $h_j=\ell_j$ and obey the same argument. □

The three factors are the latent patient query and two unit representations; the factorization uses at most r shared products. A nonzero D establishes a positive approximation gap on the eight specified latent inputs for the pairwise-additive score class. A1/A2 are nonlinear predictors whose retrained comparisons separately test explicit structural paths in the complete model.

A.1.2 Differentiating and bounding the patient-conditioned read

Fix a nonempty valid unit or relation set, finite scores and its patient query. For clarity, suppress the patient and expert indices. Let $v=\sum_k \eta_k\tanh(s_k)$ and $\eta_k=\mathrm{softmax}_k(s)$, as in Equation (3). With the normalization guard inactive, the derivative is

$$\frac{\partial v}{\partial s_j} = \eta_j\big[1-\tanh^2(s_j)+\tanh(s_j)-v\big]. \tag{18}$$

The softmax Jacobian is $\partial\eta_k/\partial s_j = \eta_k(\mathbf{1}[k=j] - \eta_j)$. Applying the product rule gives

$$\begin{aligned}\frac{\partial v}{\partial s_j} &= \sum_k \eta_k(\mathbf{1}[k=j] - \eta_j)\tanh(s_k) + \eta_j\big(1 - \tanh^2(s_j)\big) \\ &= \eta_j\big(\tanh(s_j) - v\big) + \eta_j\big(1 - \tanh^2(s_j)\big),\end{aligned}$$

which proves Equation (18). The first term redistributes attention relative to the current read. The second differentiates the item's bounded score directly. The masked softmax sets invalid weights to zero and renormalizes over valid items. Its denominator is floored at the smallest positive normal value of its floating-point type. The identities apply in real arithmetic with that floor inactive and a fixed valid set. A fixed empty set returns zero and contributes zero read gradient.

Proposition 4 (Length-independent score sensitivity)**.** *For a fixed nonempty valid set and finite real scores, with normalization guards inactive,* $v(s) = \mathrm{softmax}(s)^\top \tanh(s)$ *satisfies*

$$\|\nabla_s v\|_1 \le 3, \qquad |v(s) - v(s')| \le 3\|s - s'\|_\infty. \tag{19}$$

The constant is independent of valid-set size. With fixed unit and edge masks, fixed response scale and zero interval reads, the dedicated correction satisfies

$$|r_e - r'_e| \le \frac{3}{4}\left(\|s^U - s^{U\prime}\|_\infty + \|s^E - s^{E\prime}\|_\infty\right). \tag{20}$$

A fixed empty set contributes zero to both sides.

Proof. Both $\tanh(s_j)$ and their convex mean v lie in $[-1, 1]$. Equation (18) therefore gives

$$\left|\frac{\partial v}{\partial s_j}\right| \le \eta_j\{1 + 2\} = 3\eta_j.$$

Summation gives the gradient bound. Integrating along the segment from s to s' and applying the ℓ_1–ℓ_∞ inequality proves the perturbation bound. Finally $r_e = \kappa_e \tanh(v^U + v^E)$, with $|\kappa_e| \le 1/4$. Since $\tanh$ is 1-Lipschitz, apply the read bound to each of the two fixed sets to obtain Equation (20). An empty read is identically zero. □

These bounds control score perturbations with the valid treatment-unit and relation sets held fixed.

For the formal inputs, the date-matched interval term is zero. Write $\kappa_e = \frac{1}{4}\tanh(a_e)$, so that $r_{ie} = \kappa_e \tanh(v^U_{ie} + v^E_{ie})$. For one valid relation edge (j, k), the response chain is

$$\begin{aligned}\frac{\partial r_{ie}}{\partial s^E_{ie,jk}} &= \kappa_e\big[1 - \tanh^2(v^U_{ie} + v^E_{ie})\big]\frac{\partial v^E_{ie}}{\partial s^E_{ie,jk}} \\ &= \kappa_e\big[1 - \tanh^2(v^U_{ie} + v^E_{ie})\big]\eta^E_{ie,jk}\big[1 - \tanh^2(s^E_{ie,jk}) + \tanh(s^E_{ie,jk}) - v^E_{ie}\big].\end{aligned} \tag{21}$$

The same chain holds for a unit score with superscript U. This is the derivative of the dedicated positive-logit increment; the treatment feature also reaches the base classifier and raw router. The full interval interface and its event eligibility conditions are given in Appendix B.

Endpoint and shared-map derivatives. Fix the patient query, relation type and endpoint maps. A valid relation has distinct endpoint nodes. Its endpoint gradients are

$$\nabla_{u_j} s^E_{ie,jk} = B_{ie}u_k, \qquad \nabla_{u_k} s^E_{ie,jk} = B^\top_{ie}u_j. \tag{22}$$

To obtain them, differentiate $s^E_{ie,jk} = u_j^\top B_{ie}u_k$: $\mathrm{d}s^E_{ie,jk} = (\mathrm{d}u_j)^\top B_{ie}u_k + u_j^\top B_{ie}\mathrm{d}u_k$. Concurrent relations use symmetric B_{ie}. Independent sequential maps distinguish incoming from outgoing endpoints. A unit on several edges receives the sum of their contributions, each multiplied by the corresponding read/response chain.

For a sequential edge, let $l = W_L u_j$ and $h = W_R u_k$. Differentiating the same score with respect to the two shared maps, holding the query and endpoint features fixed, gives

$$\nabla_{W_L} s^E_{ie,jk} = r^{-1/2}(\xi_{ie} \odot h)u_j^\top, \qquad \nabla_{W_R} s^E_{ie,jk} = r^{-1/2}(\xi_{ie} \odot l)u_k^\top. \tag{23}$$

For a concurrent edge the same W_L occurs at both endpoints, so its derivative is the sum $r^{-1/2}\{(\xi_{ie} \odot W_L u_k)u_j^\top + (\xi_{ie} \odot W_L u_j)u_k^\top\}$. These score-node derivatives are multiplied by their downstream derivatives and accumulated across edges, experts, patients and supervised prefixes; the pooled relation representation supplies additional classifier and router paths. The unit nodes u_j are the trainable treatment encoder's output features. Their upstream maps combine frozen semantic inputs with recorded granularity, event-status and time metadata.

A.2 LIKELIHOOD ROUTING, EXPERT GRADIENTS AND DIRECT SUPERVISION

Direct supervision of available experts. Fix a patient, valid prefix and nonempty available-expert set. Let ζ_e be expert e's adjusted binary log-odds, so its pCR probability is $p_e = \sigma(\zeta_e)$. Its probability of the observed class is $\pi_e = p_e^y(1-p_e)^{1-y}$. The mixture assigns that class probability $\pi = \sum_e \alpha_e \pi_e$. Define likelihood responsibility as $\gamma_e = \alpha_e \pi_e / \pi$. The local classification subobjective is

$$\mathcal{L}_{\text{pred}} = -\log \pi + \sum_e \lambda_e(-\log \pi_e). \tag{24}$$

Here $\lambda_e \geq 0$ is the effective direct-classification weight after division by the positive local mixture-loss coefficient. It is fixed for the local calculation. Appendix B.6.1 defines the eligible-head, prefix and patient reductions that determine it.

First consider the mixture term. Its loss and raw-router gradient satisfy

$$\ell_{\text{mix}} = -\log \pi \leq \sum_e \alpha_e(-\log \pi_e), \qquad \frac{\partial \ell_{\text{mix}}}{\partial q_e} = \alpha_e \left(1 - \frac{\pi_e}{\pi}\right). \tag{25}$$

Proposition 5 (Mixture likelihood and routing gradient)**.** *Fix a nonempty available-expert set. Assume finite logits and learned softmax routing, with numerical floors inactive. Let $\pi_e > 0$ be expert e's probability of the observed outcome and $\pi = \sum_e \alpha_e \pi_e$. Equation* (25) *then holds. For adjusted binary expert logit difference ζ_e and pCR probability $p_e = \sigma(\zeta_e)$,*

$$\frac{\partial \ell_{\text{mix}}}{\partial \zeta_e} = \frac{\alpha_e \pi_e}{\pi}(p_e - y), \qquad \frac{\partial \text{KL}(\rho \| \alpha)}{\partial q_e} = \alpha_e - \rho_e \tag{26}$$

when other outputs and the detached target ρ are fixed.

Proof. The negative logarithm is convex and available weights sum to one, giving the likelihood inequality. Since $\partial \alpha_j / \partial q_e = \alpha_j(\mathbf{1}[j = e] - \alpha_e)$, we have $\partial \pi / \partial q_e = \alpha_e(\pi_e - \pi)$. Multiplication by $-1/\pi$ gives Equation (25). The expert likelihood derivative is $\partial \log \pi_e / \partial \zeta_e = y - p_e$, yielding the first identity in Equation (26). Finally, the detached-target KL is a constant minus $\sum_j \rho_j \log \alpha_j$; differentiating softmax cross-entropy gives $\alpha_e - \rho_e$ when its logarithm floors are inactive. □

Proposition 6 (Direct supervision of available experts)**.** *During learned routing, fix the masks and stochastic realization. With finite logits and numerical guards inactive, the output-node partial derivatives are*

$$\frac{\partial \mathcal{L}_{\text{pred}}}{\partial \zeta_e} = (\gamma_e + \lambda_e)(p_e - y), \qquad \frac{\partial \mathcal{L}_{\text{pred}}}{\partial q_e} = \alpha_e - \gamma_e. \tag{27}$$

Proof of Proposition 6. Under the stated fixed-mask, fixed-stochastic-realization and learned-routing conditions, define the likelihood responsibility $\gamma_e = \alpha_e \pi_e / \pi$. The preceding identities give $\partial \ell_{\text{mix}} / \partial \zeta_e = \gamma_e(p_e - y)$ and $\partial \ell_{\text{mix}} / \partial q_e = \alpha_e - \gamma_e$. Each direct term $\lambda_e(-\log \pi_e)$ contributes $\lambda_e(p_e - y)$ at its classifier node and zero at the distinct raw-router node. Summing these contributions gives Equation (27) for the local classification subobjective in Equation (24). Coefficients λ_e are fixed by eligibility and the implemented reductions; their relation to the full loss is detailed in Appendix B.6.1. □

The direct term is independent of the mixture responsibility. Both output-node derivatives feed shared upstream parameters, and their chain-rule contributions must therefore be combined. Jensen's bound concerns the available experts' weighted mean log loss. The resulting parameter-level signal depends on the expert error, effective supervision coefficient and upstream Jacobians.

Treatment-unit chain rule. For a valid unit u_j, write $\zeta_e = b_e + r_e$. Here b_e is the base classifier log-odds, and $r_e = 0$ for heads without a dedicated correction. Under Proposition 6's conditions,

$$\nabla_{u_j} \mathcal{L}_{\text{pred}} = \sum_{\substack{e \in \mathcal{S}(A) \\ T \in e}} \Big[(\gamma_e + \lambda_e)(p_e - y)(\nabla_{u_j} b_e + \nabla_{u_j} r_e) + (\alpha_e - \gamma_e) \nabla_{u_j} q_e \Big]. \tag{28}$$

Proof. Equation (27) gives the two output-node partial derivatives. Substituting $\zeta_e = b_e + r_e$ and applying the chain rule over the permitted expert graph yields Equation (28). Heads without treatment have zero unit-node Jacobians; heads without a dedicated response correction have $r_e = 0$. The derivatives include all incident relation-token paths. The three terms follow the pooled treatment feature into the expert classifier, the unpooled unit/relation tokens into patient-conditioned treatment reads, and the pooled feature into the raw routing score. The patient query itself uses the expert's allowed non-treatment representations, so it is fixed when differentiating a treatment-unit node. The read and endpoint derivatives in Appendix A.1 expand the dedicated-response path for the formal inputs.

Shared unit-map parameters collect these unit-node gradients through their upstream Jacobians across patients and regimens; frozen language backbones provide fixed semantic inputs. Equation (28) concerns the normalized local classification subobjective. Router KL, image attributes, load balancing and the explicit response/FFT regularizers add their own derivatives to the full objective. Gradient clipping and AdamW subsequently operate on the accumulated parameter gradients.

A.2.1 Text anchoring and the two treatment-learning paths

Let ϑ contain the shared unit-encoding and endpoint-map parameters, excluding the whole-text projection, terminal residual map R, patient-query maps and expert heads. For a record with an available whole description, write $d_\vartheta = \overline{u} + W_e\overline{e}$ and $z^T = \phi_w(w) + Rd_\vartheta$. The local objective is Equation (24). Use fixed eligibility, effective coefficients and stochastic realization, finite outputs, and inactive numerical guards. Treatment supervision is active; its three-epoch representation warm-up is considered separately.

Proposition 7 (Text anchor with a direct unit-learning path)**.** *Let g_T be the output-node derivative $\partial\mathcal{L}_{\mathrm{pred}}/\partial z^T$, holding the dedicated response nodes independent in this partial derivative. Let $\mathcal{J}$ contain the eligible IT/CT/ICT experts. Then Equation* (13) *holds and*

$$\nabla_R\mathcal{L}_{\mathrm{pred}} = g_T d_\vartheta^\top, \qquad \nabla_\vartheta\mathcal{L}_{\mathrm{pred}}\big|_{R=0} = \sum_{e\in\mathcal{J}}(\gamma_e + \lambda_e)(p_e - y)\nabla_\vartheta r_e. \tag{29}$$

Proof. The parameter block ϑ reaches the classification objective through the pooled feature z^T and the dedicated corrections r_e. Patient queries use only permitted non-treatment modalities and do not depend on this parameter block. Base classifiers and raw routers receive treatment through z^T; their contributions are included in g_T. The chain rule through the pooled feature gives $J_{d_\vartheta}^\top R^\top g_T$. Since r_e is added only to the positive-class logit, its output-node derivative is $(\gamma_e + \lambda_e)(p_e - y)$ from Equation (27). Summing the two paths proves Equation (13). Differentiating Rd_ϑ with respect to R gives $g_T d_\vartheta^\top$; the direct reads do not use R. Setting $R = 0$ eliminates only the pooled contribution to ϑ. □

The zero-initialized residual map anchors the pooled feature to the whole-text projection. Dedicated reads bypass this map and use a response-scale parameter initialized to 0.4, giving active treatment supervision a direct path to unit and relation parameters. Gradient magnitudes depend on prediction error, activation derivatives and cancellation along these paths. Equation (22) shows how each endpoint's update depends on its partner and the patient-conditioned matrix.

For unit-only records, the implementation uses $z^T = d_\vartheta$ and the pooled contribution is $J_{d_\vartheta}^\top g_T$, without R. During uniform routing, the same classifier-node identity holds with fixed mixture weights, while raw-router derivatives are bypassed. Before treatment supervision is enabled, its classification contribution is inactive. Auxiliary regularizers add their own derivatives and are not included in this local classification decomposition.

Numerical conditions. Probability accumulation uses FP32 log-softmax and log-sum-exp. Available routing weights are floored at 10^{-30} before the logarithm in the mixture computation; the CE operation normalizes the resulting two class scores. An active floor gives normalized floored mixture weights and requires differentiation through that floored computation. The detached-target router KL uses additional 10^{-8} logarithm floors. The displayed identities apply to the interior with finite logits and all relevant floors inactive. During uniform-routing warm-up, the learned raw routing score is bypassed; the learned-softmax gradient statements apply after that stage.

A.3 Proof of consistent-reindexing invariance

Proposition 8 (Treatment representation consistency)**.** *Assume deterministic evaluation and fixed whole-description semantics. Jointly reindexing units, metadata and relation endpoints leaves the structured treatment feature and patient-conditioned reads unchanged. Reversing a concurrent edge preserves its token. The sequential map retains directional capacity.*

Let $U = (u_1, \ldots, u_m)$ be valid unit tokens and let π be a permutation. Applying the same unit map to jointly permuted semantics and metadata produces $U'_j = U_{\pi^{-1}(j)}$. Replacing every edge (j, k, r) by $(\pi(j), \pi(k), r)$ preserves its token under Equation (1); any change is only in the order of edge storage. Unit and edge means therefore remain unchanged, including masked empty-set conventions. The whole-description semantic anchor is held fixed, so adding its structured residual preserves the resulting treatment feature.

For a fixed patient query, the score vector is permuted with the units. If P is the permutation matrix, then $\mathrm{softmax}(Ps) = P\,\mathrm{softmax}(s)$ and $\tanh(Ps) = P\tanh(s)$. Hence

$$(P\,\mathrm{softmax}(s))^\top P\tanh(s) = \mathrm{softmax}(s)^\top \tanh(s), \tag{30}$$

since $P^\top P = I$. The same argument applies to the masked relation read. An empty set returns zero before and after reindexing. Any interval event read also stays unchanged when endpoint, date and status metadata are reindexed consistently. Consequently, the dedicated response correction is invariant. For a concurrent edge, commutativity of elementwise multiplication gives $(W_L u_j) \odot (W_L u_k) = (W_L u_k) \odot (W_L u_j)$. Independent left/right maps allow orientation-dependent sequential values, with equality for parameter/input configurations having zero antisymmetric score. The result concerns structured-unit reindexing with whole-text semantics fixed. □

A.4 Proof of the dedicated response bound

Proposition 9 (Bounded dedicated response correction)**.** *For any valid input, $|r_{ie}| \le 1/4$. Holding the remaining expert logits fixed, the dedicated correction changes the expert's pCR probability by at most $1/16$.*

Let a and v denote the learned scale and the sum of valid reads in Equation (43). Since $|\tanh(a)| \le 1$ and $|\tanh(v)| \le 1$, $|r| \le 1/4$. If the original two logits are (ℓ_0, ℓ_1), the positive probability is $\sigma(z)$ for $z = \ell_1 - \ell_0$. Adding r only to the positive logit gives $\sigma(z + r)$. The mean-value theorem and $\sigma'(z) = \sigma(z)(1 - \sigma(z)) \le 1/4$ imply

$$|\sigma(z + r) - \sigma(z)| \le \tfrac{1}{4}|r| \le \tfrac{1}{16}. \tag{31}$$

This conditional bound isolates the dedicated response correction with all other logits fixed. Treatment features also influence each expert-specific representation and raw routing score through their separate trainable paths. □

A.5 Raw expert non-interference

Proposition 10. *For fixed evaluation parameters and availability, changing a modality outside an expert's declared subset does not change that expert's logits, response correction or raw routing score.*

Proof. Each modality encoder processes its own stream. Each private modality-subset expert has an expert-specific representation network, latent projection, expert classifier and routing head. Its representation concatenates the declared modality features; its classifier and raw routing score depend on that representation. Treatment-containing joint experts additionally use patient-conditioned treatment reads from their permitted patient modalities and, where eligible, imaging prefixes. Composition preserves independence from undeclared inputs. The final softmax denominator includes all available raw scores, coupling their normalized weights in the availability-masked probability mixture. □

A.6 Prefix consistency and single-visit identity

Proposition 11. *Assume visit preprocessing and contextual snapshots use only information eligible at the supplied prefix. Evaluation is deterministic, and adding visits does not alter earlier stage/time*

assignments, contextual snapshots or availability. Adding or perturbing future visits then cannot change an existing prefix prediction. A one-visit temporal state is exactly its imaging representation. At differentiable points,

$$z^I_{i,0} = x_{i,0}, \qquad \frac{\partial \widehat{p}_{i,t}}{\partial x_{i,s}} = 0 \quad (s > t). \tag{32}$$

Proof. Each visit is encoded independently with shared weights. Group normalization has no cross-visit running batch statistics. The history attention mask excludes future positions at every layer. By induction over layers, the state H_t depends only on positions up to t. The current-state update reads x_0, x_t, x_{t-1} and H_t, all inside that prefix. Clinical/treatment snapshots and expert availability are also prefix-restricted by assumption. Composing these operations with the private modality-subset experts preserves future non-interference. For one visit, the temporal implementation explicitly returns x_0 without a stage/time update, establishing exact identity. At differentiable points this functional independence gives Equation (32). □

The proposition establishes architectural prefix consistency under the stated input-timing assumptions. The present data use retrospective visit ordering and regimen documentation; verified delivered-treatment intervals are unavailable.

One treatment structure across imaging prefixes. Fix model parameters, treatment-unit features, relation endpoints and types, and the clinical snapshot. At deterministic evaluation, consider two eligible prefixes s and t of the same patient and an image-containing treatment expert $e \in \{IT, ICT\}$. Its fixed query map produces $\xi_{ie,s}$ and $\xi_{ie,t}$ from the respective patient states. Subtracting the two sequential-edge scores gives

$$s^E_{ie,t,jk} - s^E_{ie,s,jk} = \frac{(\xi_{ie,t} - \xi_{ie,s})^\top \left[(W_L u_j) \odot (W_R u_k)\right]}{\sqrt{r}}. \tag{33}$$

The concurrent identity replaces W_R by W_L. The observed-prefix state in Equation (41) enters the query projection in Equation (42). A change in this query updates the fixed relation token's score according to Equation (33). For the formal inputs, $v^{\text{interval}}_{ie} = 0$. The update therefore occurs through the ordinary unit/relation reads.

A.6.1 Clinical and temporal sensitivity of a shared relation

Fix treatment endpoints, model parameters and input availability. The CT query is a function of the clinical representation z^C; the ICT query also uses the image representation. Holding the other patient stream fixed, the chain rule at differentiable points gives

$$\nabla_{z^C} s^E_{jk} = r^{-1/2} J^\top_{\xi, z^C} [(W_L u_j) \odot (W_R u_k)]. \tag{34}$$

Here J_{ξ, z^C} is the Jacobian of the implemented patient query with respect to the clinical representation. The IT relation has no clinical input and its corresponding derivative is zero. This local dependency places clinical conditioning inside the relation score, before expert fusion.

For two queries ξ_s, ξ_t, the shared relation matrices satisfy

$$\|B(\xi_t) - B(\xi_s)\|_2 \le \frac{\|W_L\|_2 \|W_R\|_2}{\sqrt{r}} \|\xi_t - \xi_s\|_\infty. \tag{35}$$

Indeed, substitute the factorization of B and use submultiplicativity of the operator norm together with $\|\operatorname{diag}(\delta\xi)\|_2 = \|\delta\xi\|_\infty$. Multiplying by endpoint norms yields

$$|s^E_{t,jk} - s^E_{s,jk}| \le \frac{\|W_L\|_2 \|W_R\|_2 \|u_j\|_2 \|u_k\|_2}{\sqrt{r}} \|\xi_t - \xi_s\|_\infty. \tag{36}$$

The concurrent case again uses $W_R = W_L$. Combined with Proposition 4, this bounds the dedicated read's response to changing patient queries on fixed treatment sets. The constants depend on the learned maps and endpoint norms; no norm restriction is added to training.

The implemented temporal state updates the IT/ICT queries through available MRI histories. Clinical and treatment snapshots remain fixed across stages in the formal inputs, and additional longitudinal examinations come from I-SPY2. These identities characterize the temporal computation; the reported performance tables evaluate baseline predictions.

A.7 Conditional spectral perturbation bound

Proposition 12. *For a stage difference tensor d, its computed real band gains in $(0.5, 1.5)$ and spatially constant per-channel gate in $(0, 1)$, Equation (45) satisfies $\|\widetilde{d} - d\|_2 \le \frac{1}{2}\|d\|_2$.*

Proof. The three disjoint radial masks cover the spectrum. Their real gains define an even Fourier multiplier A with $|A(\omega) - 1| \le 1/2$ at every frequency and channel. With an orthonormal full Fourier transform F, the correction is $GF^{-1}\,\mathrm{diag}(A-1)Fd$, where G multiplies channels by their gates and $\|G\|_2 \le 1$. Parseval's identity gives

$$\|\widetilde{d} - d\|_2 \le \|G\|_2 \|\,\mathrm{diag}(A-1)\|_2 \|d\|_2 \le \tfrac{1}{2}\|d\|_2. \tag{37}$$

The real half-spectrum implementation is equivalent: non-DC and non-Nyquist last-axis entries receive conjugate multiplicity two in energy calculations. The even real radial multiplier preserves the Hermitian symmetry needed for a real inverse. □

This fixed-multiplier bound controls the stage correction with the input's computed gains and gate held fixed.

A.8 Availability-constrained probability validity

Proposition 13. *If at least one modality is available and every available expert outputs a valid binary probability, Equation (11) produces a valid probability in $[0, 1]$ and places zero weight on unavailable experts.*

Proof. Each available raw score has a positive exponential, and the denominator is the positive sum of those exponentials. Thus available weights are nonnegative and sum to one; masked experts receive zero. A convex combination of values in $[0, 1]$ remains in $[0, 1]$. □

B Implementation and optimization details

B.1 Cohort records and input coverage

The study includes 1,522 input records across four cohorts; 1,514 have evaluable reference endpoints (Table 4). I-SPY2 and Duke form the US development group. Hospital-A and Hospital-B are anonymized hospitals in different regions of a second country. All study data and reference annotations were reviewed by physicians. Physicians at Hospital-A and Hospital-B collected and annotated their respective datasets. Appendix G documents cohort-specific endpoint sources and coding rules.

Hospital-B endpoints use post-treatment pathology exclusively. Absence of residual invasive breast carcinoma with negative regional lymph nodes defines pCR. Residual invasive breast carcinoma or regional nodal positivity defines non-pCR. Eight records with incomplete pathology receive predictions but are excluded from scoring and threshold fitting.

Treatment records span named medications, classes, alternatives and regimen-level descriptions. I-SPY2 + Duke, Hospital-A and Hospital-B contain 2,504, 315 and 74 unit occurrences, respectively (Table 4b). Source inputs include 368 alternative-set units and 184 sequential edges, with no explicit concurrent edges. Hospital-A has 76 concurrent edges; Hospital-B has 29 regimen-level units. The same treatment encoder processes these different record structures.

Development and Hospital-A have baseline masks and available clinical and treatment branches. Hospital-B has no tumor masks. Under the fixed input adapter, 195 of its 197 inputs have usable clinical representations and 38 have treatment representations. Among the 189 scored records, these counts are 187 and 35. Treatment is unavailable for 154 scored records and 159 of all 197 inputs. Two records contain complex clinical reports but yield no usable fields under the adapter's aggregation rules (Appendix D). Supplementary Figure S1 displays input coverage.

Additional examinations are available for 582 I-SPY2 patients and enter source training as observed histories. Duke and both hospitals contribute baseline inputs. Performance tables evaluate baseline MRI with retrospective clinical and treatment snapshots. Their availability at the pretreatment decision point was not verified. Treatment availability is associated with outcome.

Table 4: **Study records and treatment-input composition.** (a) Four cohort counts; performance for the two US sources is reported as I-SPY2 + Duke. (b) Baseline treatment-unit occurrences and explicit relation-edge counts over all input records in each group. Unit counts include repeated occurrences across records.

(a) Cohort records and reference outcomes

Cohort	Input	Scored	pCR	non-pCR	Pending
I-SPY2	980	980	316	664	0
Duke	249	249	25	224	0
Hospital-A	96	96	43	53	0
Hospital-B	197	189	76	113	8
Total	1,522	1,514	460	1,054	8

(b) Treatment-unit occurrences and explicit relation edges

Input group	Named	Class	Alternatives	Regimen	Concurrent	Sequential
I-SPY2 + Duke	2,071	62	368	3	0	184
Hospital-A	283	5	2	25	76	39
Hospital-B	39	0	6	29	0	9

B.2 Inputs and representation dimensions

For patient record i and observed prefix t, let $X_{i,\le t}$ denote MRI and $C_{i,t}$ denote clinical context. The treatment input consists of whole-description text $w_{i,t}$, supported units $\mathcal{U}_{i,t}$ and typed relations $\mathcal{E}_{i,t}$. With availability $A_{i,t}$, the formal target is

$$p_\theta(Y_i = 1 \mid X_{i,\le t}, C_{i,t}, w_{i,t}, \mathcal{U}_{i,t}, \mathcal{E}_{i,t}, A_{i,t}). \tag{38}$$

Here Y_i is the patient-level pCR endpoint.

The L3 runs use cached paired MRI tensors of shape $2 \times 88 \times 176 \times 176$ with target spacing 2 mm on each axis. Both phases and the mask receive aligned spatial processing on derived training tensors. Shared-phase stage strides are $(2, 4, 4)$ followed by three $(2, 2, 2)$ stages. Modality features and unit preprojections have dimension 128. Fixed clinical slots and typed field values each have dimension 16. Expert latent dimension and pair rank are both 16. The history transformer has one layer, four heads and feed-forward width 256. Regular dropout is 0.15. Five visit-stage indices include missing/unspecified handling; reported performance concerns baseline prediction.

Both language backbones provide 768-dimensional semantic features, with a token limit of 512 and up to 32 treatment units per record. Observed text is encoded in local chunks with padding-masked pooling and permanently frozen evaluation-mode backbones. The whole clinical semantic residual is 0.1. Each source training partition determines the active field set. Numeric/categorical encodings use fixed definitions; explicitly observed ER/PR and clinician-reviewed wording are preserved as recorded.

The *Implementation and Scoring Guide* identifies the retained language-model weights, tokenizer assets and configuration files by SHA-256, together with the five-fold code and configuration fingerprints.

The model therefore has 220,998,261 total parameters, approximately 3.21M registered trainable parameters and 217.79M frozen parameters. The 4,096 shared endpoint-map weights are included in the 240,640-parameter treatment encoder. This ledger includes the inactive date-matched interval-query interfaces.

For field f, let b_f be frozen BioBERT semantics, k_f indicate known field recency and $\bar{d}_f$ be its nonnegative recency when known and zero otherwise. The trainable recency map forms

$$\widetilde{b}_f = b_f + \tanh\big(W_r[\log(1 + \bar{d}_f), k_f] + a_r\big)\,. \tag{39}$$

In the formal inputs all recencies are unknown, so the recency input is $(0, 0)$ and the learned offset $\tanh(a_r)$ is retained. For observed-field indicator m_f, typed-value indicator o_f and typed vector v_f, the clinical slot and resulting feature are

$$c_f = m_f\big(o_f W_v v_f + 0.1 W_s \widetilde{b}_f + 0.05 e_f\big), \qquad z^C = \phi_C([c_1; \ldots; c_{32}]) + 0.1\phi_0(b_{\text{whole}}). \tag{40}$$

Table 5: **Registered parameter ledger.** Counts follow registered tensor elements and frozen status in the checkpoint-matched model configuration. Registered trainability includes conditional modules; the parameters receiving gradients depend on input availability and training stage.

Module	Trainable	Frozen
Shared-phase MRI encoder	2,003,129	0
Current-visit history encoder	266,752	0
Whole clinical projection	99,968	0
Fixed clinical fields	95,232	0
Clinical recency map	2,304	0
Treatment-unit and relation encoder	240,640	0
Seven private modality-subset experts	497,208	0
Image HR/HER2 attribute heads	516	0
Frozen BioBERT backbone	0	108,310,272
Frozen PubMedBERT backbone	0	109,482,240
Total	3,205,749	217,792,512

Each missing slot is zero; availability masks additionally set an absent field-branch or whole-description output to zero. The supported fields are age, HR, HER2, subtype, menopause and observed ER/PR. Field dropout also disables the whole-description bypass. Slot presence weights report uniform allocation over the observed fields.

Across examinations, stage embeddings and known-time inputs enter a causal history transformer. A baseline-only record follows $z_0 = x_0$ exactly. For later visits,

$$z_t = x_t + g_t \odot \psi([x_0, x_t, x_t - x_0, x_t - x_{t-1}, H_t]), \qquad t > 0, \tag{41}$$

where H_t is restricted to the observed prefix and g_t is a channel gate computed from x_t and H_t. P0/P1 are within-examination phases, whereas this prefix indexes separate examinations. The inputs reuse clinical and treatment snapshots across stages; changes across prefixes come from MRI, stage and available relative examination times.

B.3 Treatment evidence and operational granularity

Patient query and structured residual. The patient-query projection is expert-specific. It applies layer normalization to the permitted non-treatment representation, followed by an affine projection and a hyperbolic tangent:

$$\xi_{ie} = \tanh\big(W_e^q \,\mathrm{LN}(z_{i,e\setminus T}) + b_e^q\big) \in \mathbb{R}^r. \tag{42}$$

Here $z_{i,e\setminus T}$ is the image representation for IT, the clinical representation for CT, and their concatenation for ICT. The input dimension is 128 for IT and CT and 256 for ICT; the output dimension is $r = 16$. These inputs contain no treatment features, providing the fixed-query dependency used in the treatment-unit derivatives.

The structured residual $\mathcal{R}$ in Equation (2) is a bias-free 128×128 linear map initialized to zero. With whole-description semantics available, its initial structured increment into the pooled classification/routing feature is zero. The dedicated patient-conditioned reads use the unit and relation tokens before pooling. When only units are available, the treatment feature is $\overline{u} + W_e\overline{e}$, without $\mathcal{R}$. Empty unit and relation means are zero.

The unit and relation reads contribute a bounded correction to the expert's positive-class logit:

$$r_{ie} = \tfrac{1}{4}\tanh(a_e)\tanh(v_{ie}^U + v_{ie}^E). \tag{43}$$

The learned scalar a_e controls the correction's scale. Both the dedicated reads and the pooled treatment feature receive response supervision.

Record-supported input construction. The unit states distinguish named drug/item, drug class, alternatives, unresolved regimen and unknown granularity. Event status distinguishes delivered, planned and unspecified; known administration dates carry a separate indicator. The parser first uses a recorded arm-component list when present, splitting "+" components without creating relation edges.

Otherwise, directed edges require an initial "Sequential" description with "followed by" or "then" separating the two component blocks. An initial "Concurrent" list creates symmetric edges only between named-item or drug-class endpoints. Alternative and unresolved blocks retain their recorded granularity. Associated mechanism descriptions are retained from the record, and all ingredients remain grounded in documented treatment.

Unit granularities and relation types are parser-derived model inputs, summarized in Table 4 and Supplementary Figure S1. Their frequencies characterize cohort-specific treatment documentation. The present inputs have zero dated, confirmed delivered units, leaving interval reads inactive. The pooled source baseline inputs contain 184 sequential and zero concurrent edges; the concurrent interface shares its endpoint map with the sequential encoder.

Date-matched interval interface. The implementation retains the full response expression

$$r_{ie} = \tfrac{1}{4}\tanh(a_e)\tanh(v_{ie}^{U} + v_{ie}^{E} + v_{ie}^{\text{interval}}), \qquad v_{ie}^{\text{interval}} = 0 \text{ for the formal inputs.} \tag{44}$$

Only IT and ICT have an image-interval query. It requires at least two valid examinations with known chronological times and a positive interval. Eligible treatment events have known dates and delivered status. A unit belongs to an interval when its event day lies strictly after its start and at or before its end. A pair requires both endpoints to be eligible and to occur after the first supplied examination. The later endpoint day determines its interval assignment. The interval query reads the two examination representations and their signed difference, together with the permitted patient query. Matched unit and relation scores are transformed by $\tanh$ and averaged separately, with empty matches returning zero. CT has no interval branch. For the study inputs, the interval-eligibility set is empty and Equation (44) reduces to Equation (43). Unknown-date and unspecified-status embeddings remain part of the ordinary unit representation.

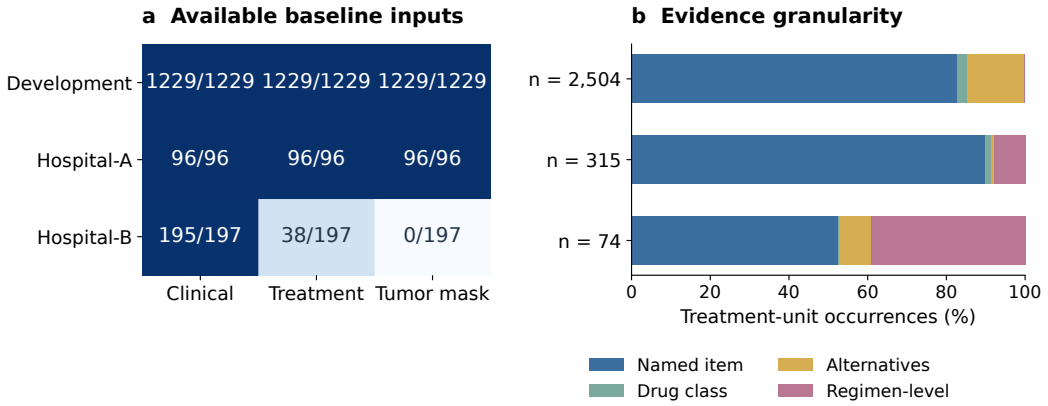


Figure S1: **Cross-setting input and treatment heterogeneity.** (a) Clinical/treatment branch and tumor-mask availability, using full input denominators. Individual fields may be missing within an available branch. (b) Recorded treatment-unit composition in the same cohort order. A common encoder preserves named items, classes, alternatives and regimen-level descriptions. Counts characterize the study inputs.

B.4 Recorded regimens and their learned representations

Table 6 traces four records through the input adapter: named constituents, grouped expressions, explicit concurrent relations and an unresolved abbreviation. Component lists supply units; explicit relation syntax supplies edges. Here alternative-set names a text-parsing category for slash-delimited expressions; the source record supplies the clinical administration context.

A regimen-level view across cohorts. Among the 1,514 records with resolved reference outcomes, 1,360 have treatment input. Deduplication uses complete text, ordered unit texts, granularity, date/status metadata and typed relations. This yields 39 unique signatures: 13 in I-SPY2, three in Duke, nine in Hospital-A and 17 in Hospital-B, with three shared by both hospitals. Figure S2 projects each available-treatment signature once, using treatment features alone.

Table 6: **Examples of recorded treatment structure.** The displayed fragments and unit labels summarize existing inputs without patient identifiers. "No explicit edge" denotes the constructed relation set. The complete recorded descriptions remain the semantic input.

Cohort	Recorded expression	Unit labels and retained granularity	Relation structure
I-SPY2	Paclitaxel combined with Neratinib	Paclitaxel; Neratinib. Two named items from the recorded arm.	No explicit edge
Duke	Sequential Anthracycline/Cyclophosphamide followed by a Taxane (Paclitaxel/Docetaxel)	Two grouped expressions, each retained under the parser's alternative-set type.	One directed sequential edge
Hospital-A	Concurrent Docetaxel (Taxane), Doxorubicin (Anthracycline), and Cyclophosphamide	Docetaxel; Doxorubicin; Cyclophosphamide. Three named items.	Three concurrent edges
Hospital-B	Recorded neoadjuvant therapy regimen: EC-T	The recorded EC-T description remains one regimen-level unit.	No explicit edge

Feature extraction and projection. The left panel uses the frozen PubMedBERT last-hidden-state masked mean before learned projection (768 dimensions). The right panel uses the complete pure-treatment feature after whole-description projection and the pooled unit–relation residual (128 dimensions). We preselect the fold-0 best checkpoint for this illustration and verify that its frozen language-model parameters exactly match the local backbone. Both panels use the same 39 signatures. Model evaluation is deterministic, dropout is inactive, and no parameter is updated. None of the 84 unique encoded whole-description or unit strings exceeds the saved 512-token limit. The source checkpoint and input fingerprints are retained in the accompanying aggregate analysis metadata.

Each panel uses L2-normalized features and exact t-SNE with cosine distance, perplexity 5, random initialization, seed 42, early exaggeration 12, automatic learning rate and at most 2,000 iterations. The panels have independently fitted axes. They compare frozen semantics and learned treatment representations, complementing the predictive comparison and controlled-retraining tables.

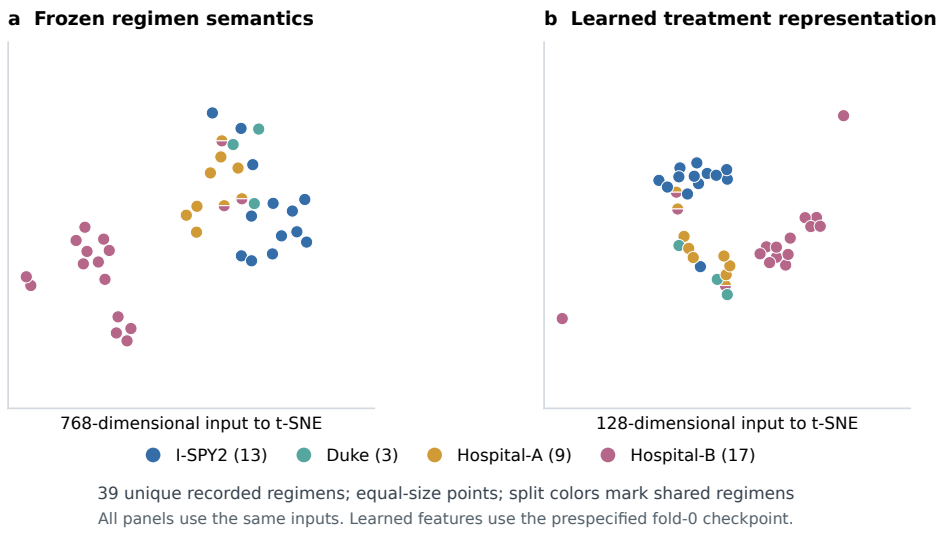


Figure S2: **Recorded regimen representations across cohorts.** Each equal-size point represents one of 39 unique treatment input signatures from endpoint-resolved records. (a) Frozen whole-description semantics. (b) Learned treatment features from the prespecified fold-0 checkpoint. Colors indicate cohort membership; split colors mark three shared hospital signatures. Legend counts refer to unique signatures. Both panels use the same treatment inputs and fixed projection settings.

B.5 FREQUENCY STATISTICS

The shared residual stages have widths 24, 48, 96 and 160. At scale ℓ, the signed feature difference is $d_\ell = E_\ell(P1) - E_\ell(P0)$. Both phase streams continue through the shared encoder; frequency modulation acts on their difference in a side branch. Real three-dimensional transforms give

$$\widetilde{d}_\ell = d_\ell + g_\ell \odot \left[\mathcal{F}^{-1}\{A_\ell \odot \mathcal{F}(d_\ell)\} - d_\ell\right], \quad A_{\ell cb} = 1 + \tfrac{1}{2}\tanh(\beta_{\ell cb} + h_{\ell cb}). \tag{45}$$

The real gains are constant within each radial band and vary by channel and patient; $g_\ell \in (0, 1)$ is spatially constant within a channel. Occupancy-based ROI/background descriptors of both phase streams and their modulated difference are pooled at all four scales. Raw phase statistics and deepest-scale attention pooling complete the visit representation. Missing or empty mask regions select global-statistic fallback.

At each stage, normalized radius uses the Euclidean norm of frequencies measured relative to axis Nyquist frequencies, divided by $\sqrt{3}$. Radial boundaries 0.25 and 0.60 define low, middle and high bands. For half-spectrum Z and conjugate multiplicity $m(\omega)$, the channel-averaged band power is

$$P_b = \frac{1}{C}\sum_c \sum_\omega m(\omega) M_b(\omega) |Z_c(\omega)|^2, \quad L_b = \log\left(1 + \frac{P_b}{n_b}\right), \quad R_b = \frac{P_b}{\sum_{b'} P_{b'} + \epsilon}, \tag{46}$$

where $n_b = \sum_\omega m(\omega) M_b(\omega)$ is clamped to at least one. The controller has 12 inputs. Six are the three L_b and three R_b values. The others are ROI/background means, ROI/background log mean-squared features, mean spatial standard deviation and spectral RMS. The gate replaces the final RMS input with mean filtered-feature standard deviation. Controllers and FFT/power calculations use FP32. All four stages are active, and gains receive a squared departure-from-one penalty.

B.6 PATIENT-NORMALIZED SUPERVISION

For a valid supervised prefix loss ℓ_{it}, define

$$\mathcal{P}_i(\ell) = \begin{cases} \frac{1}{2}\ell_{i0} + \frac{1}{2|F_i|}\sum_{t\in F_i} \ell_{it}, & |F_i| > 0 \text{ and baseline valid}, \\ \frac{1}{|V_i|}\sum_{t\in V_i} \ell_{it}, & \text{otherwise}, \end{cases} \tag{47}$$

where F_i and V_i are valid follow-up and supervised prefix sets. Valid supervision defines each reduction. At each prefix, an auxiliary group averages the available experts' losses. Source cohort-by-label weights are $w_s = N_{\mathrm{tr}}/(K_s n_s)$ for the K_s observed training strata. These fixed weights multiply patient-normalized response losses directly and retain their contribution across microbatches.

After the image/clinical warm-up, the complete joint objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\mathrm{mix}} + 0.1\mathcal{L}_{\mathrm{uni}} + 0.3\mathcal{L}_{\mathrm{bi}} + 0.1\mathcal{L}_{\mathrm{ICT}} + 0.5\mathcal{L}_D + 0.5\mathcal{L}_I + 0.05\mathcal{L}_{\mathrm{attr}} \\ + 0.001\mathcal{L}_{\mathrm{FFT}} + 0.001\mathcal{L}_{\mathrm{balance}} + 0.0001\mathcal{L}_{\mathrm{response}} + 0.05\mathcal{L}_{\mathrm{route}}.\end{aligned} \tag{48}$$

Here D is the mean of available I and C losses. The unimodal and bimodal terms first average available experts within their respective groups; ICT and I receive the additional direct CE terms shown. Supervised terms use the valid-prefix reduction over their own eligible heads and prefixes. Image attribute supervision uses observed HR and HER2; its group means are patient-normalized with unit stratum weights. FFT penalties use ordinary valid-prefix means within patients and then across patients. The response penalty averages squared increments in available IT, CT and ICT branches with the supervised group reduction and unit stratum weights. Unresolved pCR endpoints are excluded throughout.

For a training patient/prefix, let c_e be its available expert cross-entropy. The source-training-only target $\rho_e = \mathrm{softmax}_e(-\mathrm{sg}(c_e))$ is detached, and routing minimizes $\mathrm{KL}(\rho\|\alpha)$. Weak load balancing softly compares microbatch mean routing mass with mean availability-allowed uniform mass. It uses ordinary within-patient prefix means and operates when at least two labeled patients are present.

During epochs 1–3, direct image loss has weight 1.0 and the available I/C mean has weight 0.5; image attributes and the FFT constraint remain active. Epochs 4–8 train all expert losses with uniform available routing. Learned routing, epoch selection, scheduling and early stopping begin in epoch 9. This progression establishes constituent predictions before patient-dependent routing competition.

B.6.1 Effective coefficients in the local prediction objective

Fix a physical microbatch, its observed/dropout masks and a learned-routing step. The five directly supervised head groups are

$$G_{\text{uni}} = \{I, C, T\}, \quad G_{\text{bi}} = \{IC, IT, CT\}, \quad G_{\text{ICT}} = \{ICT\}, \quad G_D = \{I, C\}, \quad G_I = \{I\},$$

with coefficients $c_G = (0.1, 0.3, 0.1, 0.5, 0.5)$ in that order. Let S_{it} be the available-expert set for a labeled, valid prefix. The count $h_{it}^G = |G \cap S_{it}|$ gives the available heads in group G. Let N_G count microbatch patients with at least one valid prefix for which $h_{it}^G > 0$. The weight ω_{it}^G is the patient-normalized baseline/follow-up weight over that group's own valid prefixes. The auxiliary coefficient multiplying expert e's CE at this patient and prefix is

$$a_{ite} = \sum_G c_G \frac{w_i \omega_{it}^G}{N_G h_{it}^G} \mathbf{1}[e \in G \cap S_{it}]. \tag{49}$$

Terms with no eligible head or patient are defined as zero. The main mixture coefficient is

$$m_{it} = \frac{w_i \omega_{it}^{\text{mix}}}{N_{\text{mix}}} > 0, \qquad \lambda_{ite} = \frac{a_{ite}}{m_{it}}, \tag{50}$$

where N_{mix} and ω_{it}^{mix} use the main supervision mask. Dividing this prefix's classification contribution by m_{it} gives Equation (24); its output-node gradients in the unnormalized loss are m_{it} times Equation (27). The notation λ_e in the local statement suppresses the fixed patient and prefix indices.

When every head is available for every microbatch patient/prefix, all group patient counts and prefix weights coincide. In this particular case,

$$(\lambda_I, \lambda_C, \lambda_T, \lambda_{IC}, \lambda_{IT}, \lambda_{CT}, \lambda_{ICT}) = \left(\frac{47}{60}, \frac{17}{60}, \frac{1}{30}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right).$$

Missing modalities and modality dropout determine each sample's eligible-head, prefix and patient denominators; the simplified coefficients above apply when those denominators coincide. Source weights are fixed from the full training partition and reused across microbatches. Gradient accumulation multiplies each microbatch by its patient count divided by the accumulation-window patient count; this common factor changes the gradient scale while preserving λ_{ite}. Epochs 1–3 have zero mixture coefficient, so the positive-coefficient normalization starts in joint training. Epochs 4–8 bypass learned routing; Proposition 6 concerns the subsequent learned-routing phase.

B.6.2 Source-stratum identity for the primary mixture risk

Consider the full source-training population of N labeled patients, each having a valid primary prediction path. Let $L_i^P = \sum_t \omega_{it} \ell_{it}$ be its primary mixture CE after patient-normalized prefix reduction, with $\sum_t \omega_{it} = 1$. For source-cohort–outcome stratum $g(i)$ of size $n_{g(i)}$ and K_s nonempty strata, the weight is $w_i = N/(K_s n_{g(i)})$. The patient-reduced risk is

$$\widehat{R} = \frac{1}{N} \sum_i \frac{N}{K_s n_{g(i)}} L_i^P = \frac{1}{K_s} \sum_{g=1}^{K_s} \frac{1}{n_g} \sum_{i: g(i)=g} L_i^P. \tag{51}$$

Grouping by source stratum and expanding the patient reduction yields

$$\begin{aligned} \widehat{R}_{\text{mix}} &= \frac{1}{N} \sum_{i=1}^{N} \frac{N}{K_s n_{g(i)}} L_i^P \\ &= \frac{1}{K_s} \sum_{g=1}^{K_s} \frac{1}{n_g} \sum_{i: g(i)=g} \sum_t \omega_{it} \ell_{it}. \end{aligned} \tag{52}$$

This proves Equation (51): source-stratum weighting assigns equal total weight to each source-cohort–outcome stratum in the primary empirical risk. Patients within a stratum have equal total coefficient, and visit weights partition each patient's mass. The four strata are I-SPY2/pCR, I-SPY2/non-pCR, Duke/pCR and Duke/non-pCR, crossing the two US source cohorts with the two outcomes. Physical microbatches are sampled pieces of this primary risk. The identity is restricted to full-coverage primary supervision. Auxiliary groups use their own eligible-patient normalization N_G, with stratum mass determined by head availability; each regularizer retains its separate reduction.

B.7 Training settings and fold integrity

Training uses AdamW with initial learning rate 10^{-4} and weight decay 10^{-3}, physical batch size two and eight patient-normalized accumulation steps. The nominal effective batch is 16; the final partial group uses its actual patient count. Learning-rate scheduling uses a plateau patience of four, reduction factor 0.5 and minimum 10^{-6}. Training uses FP32, maximum 100 epochs and patience 15. Six data-loader workers, image chunks of one, language chunks of eight and gradient checkpointing manage memory. Seeds are 42 for partitions and $42 + k$ for fold k initialization; deterministic-kernel enforcement is disabled.

Training applies patient-shared mask dropout 0.5, treatment dropout 0.4, modality dropout 0.15 and clinical field dropout 0.2 while retaining a valid prediction path. Phase- and visit-aligned geometric/intensity augmentation acts on derived tensors and preserves the source records. Every source patient's visits remain in one outer/inner partition. Source-inner selection and thresholds are locked before outer and hospital prediction.

Inner selection objective. Within each cohort, patient and prediction task, the latest available prefix supplies one inner-validation score. Cohort-by-task AUC cells containing both classes receive reliability weight $\sqrt{n_+ n_- / (n_+ + n_-)}$. Weighted cell means are computed separately for the baseline and non-baseline task blocks; the nonempty block means are then averaged equally. If no cell contains both classes, the selection score is negative inner patient-mean CE. Beginning in epoch nine, a checkpoint improves the best score when it exceeds it by more than 10^{-12}; scores within that tolerance are resolved by lower inner patient-mean CE. Checkpoint selection uses these source-inner scores before outer and hospital prediction.

C Statistical interpretation and predictive values

C.1 Six metrics and population dependence

With pCR positive, $N = TP + FN + TN + FP$,

$$\mathrm{ACC} = \frac{TP + TN}{N}, \qquad \mathrm{SEN} = \frac{TP}{TP + FN}, \qquad \mathrm{SPE} = \frac{TN}{TN + FP},$$
$$\mathrm{PPV} = \frac{TP}{TP + FP}, \qquad \mathrm{NPV} = \frac{TN}{TN + FN}. \tag{53}$$

A zero denominator is undefined. AUC uses positive–negative score comparisons with half credit for ties. Hospital ensembling precedes AUC and threshold fitting; each record contributes one ensemble score and one reference endpoint.

Writing $\pi = P(Y = 1)$, Bayes' rule gives

$$\mathrm{PPV} = \frac{\pi\,\mathrm{SEN}}{\pi\,\mathrm{SEN} + (1-\pi)(1-\mathrm{SPE})}, \qquad \mathrm{NPV} = \frac{(1-\pi)\,\mathrm{SPE}}{(1-\pi)\,\mathrm{SPE} + \pi(1-\mathrm{SEN})}. \tag{54}$$

These identities express the prevalence dependence of predictive values. Each cohort's PPV and NPV are calculated from its own outcomes and decisions. Source pooled AUC includes both within- and between-cohort positive–negative comparisons and describes discrimination in the combined development population.

Youden selection maximizes sensitivity plus specificity minus one. For Hospital-B local folds $B_1, \ldots, B_5$ and equal-weight ensemble probability $\bar{p}_i$,

$$\tau_{-j} \in \arg\max_{\tau} \{\mathrm{SEN}_{B \setminus B_j}(\tau) + \mathrm{SPE}_{B \setminus B_j}(\tau) - 1\}, \qquad \widehat{y}_i = \mathbf{1}[\bar{p}_i \geq \tau_{-j}], \quad i \in B_j. \tag{55}$$

Source-inner fitting retains the highest threshold among values within absolute tolerance 10^{-12} of the maximal Youden value. Hospital-B fitting retains the first exact maximal-Youden threshold in the ROC ordering. Both retain all ROC candidates. An infinite all-negative candidate is represented by a finite cutoff above the largest score. The formal source scorer uses a float32-safe increment; Hospital-B uses a float64 increment. Cross-fitting uses each local record for cutoff fitting only when that record is outside the corresponding held-out threshold fold. Hospital-B classification is therefore locally label-adapted, with source model parameters frozen.

The bootstrap intervals condition on frozen probabilities and fixed decisions. Training and threshold fitting are not repeated. The controlled ablation analysis uses 1,000 class-stratified record resamples with seed 42. Each paired AUC difference uses identical sampled indices for both variants. Its 95% interval is the 2.5th–97.5th percentile range of those differences. The intervals are reported individually without multiplicity adjustment. PPV is undefined whenever TP+FP=0, including L1 in Hospital-A. All Hospital-B entries in Table 3 have positive PPV denominators.

D FROZEN EXPERT RESTRICTIONS

D.1 EXPERT RESTRICTIONS AND EXACT MIXTURE ALGEBRA

These inference-time restrictions quantify information dependence with all model parameters fixed, including the encoders, expert classifiers and raw routing heads. Removing I excludes I, IC, IT and ICT, leaving C, T and CT when available. Removing C leaves I, T and IT; removing T leaves I, C and IC. A single-modality restriction retains its corresponding private modality-subset expert. These retained sets contain precisely the heads supported by the selected modalities. Source predictions use each patient's outer model. Each hospital model is restricted separately before its probabilities are averaged.

Let R_i contain the retained experts whose inputs are available. Renormalizing their original weights gives

$$p_i^R = \frac{\sum_{e \in R_i} \alpha_{ie} p_{ie}}{\sum_{e \in R_i} \alpha_{ie}}. \tag{56}$$

Proposition 14 (Restricted-expert renormalization)**.** *Let S_i be the original available expert set and $R_i \subseteq S_i$ a nonempty retained set. If retained experts' inputs are unchanged, their restricted mixture is exactly Equation* (56)*, with the original weights renormalized over R_i.*

Proof. Private modality-subset experts preserve q_{ie} and p_{ie} for $e \in R_i$. Write $Z_S = \sum_{e \in S_i} \exp(q_{ie})$ and $Z_R = \sum_{e \in R_i} \exp(q_{ie})$. Since $\alpha_{ie} = \exp(q_{ie})/Z_S$, dividing each retained weight by $\sum_{e \in R_i} \alpha_{ie} = Z_R/Z_S$ gives $\exp(q_{ie})/Z_R$, precisely the available-expert softmax on R_i. Substitution gives Equation (56). □

This identity supports recomputation from stored subset-head probabilities and weights. Full-mixture reconstruction matches the model predictions within 1.02×10^{-7}. Restriction precedes hospital ensembling because patient-dependent renormalization and averaging generally do not commute.

Pointwise consequence. For one frozen model and patient, partition the available heads into a nonempty retained set R and removed set D. Under Equation (11), keep the retained inputs and output scores fixed and let $\beta = \sum_{e \in D} \alpha_e$. If $0 < \beta < 1$, define p_R and p_D by within-set renormalization of the original weights. Then

$$p = (1 - \beta) p_R + \beta p_D, \qquad |p - p_R| = \beta |p_D - p_R| \leq \beta. \tag{57}$$

The decomposition follows by splitting the original sum. Both group predictions lie in $[0, 1]$, establishing the bound. When D is empty, the difference is zero without defining an empty-set p_D. For an equal-weight ensemble of M models, each restricted before averaging, the triangle inequality gives

$$|\bar{p} - \bar{p}_R| \leq \frac{1}{M} \sum_{m=1}^{M} \beta^{(m)}. \tag{58}$$

Consequently a positive–negative pair keeps its original score ordering if its original positive margin exceeds the sum of its two pointwise bounds. This sufficient condition concerns that pair under the defined frozen restriction. Aggregate AUC is evaluated over all eligible positive–negative pairs in the results below.

D.2 SIX-METRIC RESULTS WITH EXPLICIT ELIGIBLE POPULATIONS

Each restriction receives source-inner cohort-specific cutoffs fitted to its frozen inner predictions. The restriction analysis uses the first exact maximum in ROC order for these diagnostic source cutoffs,

distinct from the source scorer's tolerance-based highest-cutoff rule. Hospital-A uses 0.5. Hospital-B preserves the five partitions defined on the original 189 scored records and filters fitting and held-out subsets within those partitions for each restriction. Eligibility requires at least one retained modality: Hospital-B has $N = 187$ for without-I/C-only, $N = 35$ for T-only and $N = 189$ otherwise. Source and Hospital-A retain their full denominators.

Table 7: **Frozen expert restrictions: combined development.** L3 experts are evaluated under restricted availability with restriction-specific source-inner cutoffs. AUC is a percentage; other metrics are percentages.

Configuration	N	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
Full	1229	71.12% (67.93%–74.26%)	65.58 (62.89–68.35)	70.09 (65.40–74.78)	63.85 (60.59–67.12)	42.68 (40.03–45.65)	84.75 (82.57–86.92)
Without I	1229	71.77% (68.73%–74.82%)	66.15 (63.55–68.84)	67.16 (62.46–71.85)	65.77 (62.95–68.92)	42.96 (40.22–45.94)	83.91 (81.94–86.12)
Without C	1229	65.04% (61.85%–68.33%)	60.29 (57.77–63.30)	43.40 (37.83–49.27)	66.78 (63.74–70.05)	33.41 (30.05–37.10)	75.45 (73.57–77.63)
Without T	1229	69.61% (66.51%–72.80%)	63.71 (61.02–66.40)	62.76 (57.48–68.04)	64.08 (60.70–67.23)	40.15 (37.23–43.00)	81.75 (79.52–83.95)
I only	1229	56.18% (52.86%–59.66%)	55.49 (52.97–58.26)	33.14 (28.44–38.13)	64.08 (60.81–67.23)	26.16 (22.86–29.59)	71.39 (69.60–73.26)
C only	1229	69.69% (66.55%–72.75%)	63.87 (61.19–66.56)	56.30 (51.32–61.58)	66.78 (63.62–69.71)	39.43 (36.36–42.51)	79.92 (78.03–81.92)
T only	1229	66.30% (63.15%–69.56%)	62.33 (59.72–65.09)	56.01 (50.73–61.29)	64.75 (61.60–67.91)	37.90 (34.91–41.07)	79.31 (77.25–81.47)

Table 8: **Frozen expert restrictions: Hospital-A.** Each of five source models is restricted first; probabilities are then averaged and classified at 0.5. N/A denotes an undefined denominator.

Configuration	N	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
Full	96	83.68% (75.12%–91.18%)	76.04 (67.71–84.38)	79.07 (67.44–90.70)	73.58 (62.22–84.91)	70.83 (61.11–81.40)	81.25 (72.92–90.91)
Without I	96	85.26% (77.18%–92.19%)	77.08 (68.75–85.42)	86.05 (76.69–95.35)	69.81 (56.60–83.02)	69.81 (61.11–80.00)	86.05 (76.74–95.12)
Without C	96	77.31% (66.60%–86.97%)	72.92 (63.54–81.25)	65.12 (48.84–79.07)	79.25 (67.92–88.68)	71.79 (60.00–83.87)	73.68 (65.45–82.61)
Without T	96	79.42% (69.85%–88.24%)	71.88 (62.50–80.21)	60.47 (46.51–74.42)	81.13 (71.70–92.45)	72.22 (59.99–86.22)	71.67 (63.49–79.66)
I only	96	48.66% (36.99%–60.21%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
C only	96	82.32% (73.54%–89.82%)	68.75 (59.38–78.12)	86.05 (74.42–95.35)	54.72 (41.51–67.92)	60.66 (53.33–69.23)	82.86 (71.43–93.94)
T only	96	79.55% (69.68%–87.89%)	76.04 (66.67–84.38)	72.09 (58.14–86.05)	79.25 (67.92–88.68)	73.81 (63.04–85.30)	77.78 (69.64–87.04)

On the same 187 Hospital-B records eligible without I, full/without-I AUC is 80.07%/80.92%. On 35 T-available records, full/T-only AUC is 90.91%/72.35%. These comparisons hold the scoring population fixed. Treatment availability is outcome-associated: the available group contains 24 pCR and 11 non-pCR records, and the unavailable group contains 52 pCR and 102 non-pCR records. Removing the treatment pathway jointly changes access to documented semantics and its routing context.

The full predictor additionally covers two Hospital-B records that have no usable clinical or treatment representation under the retained adapters, giving 189 scored records versus 187 for C/T/CT. Their

Table 9: **Frozen expert restrictions: Hospital-B.** Five-model probability means and local cross-fitted Youden thresholds. Each row displays its eligible N. T-only PPV of 100% is based on 11 positive predictions among 35 eligible records.

Configuration	N	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
Full	189	79.56% (72.89%–85.78%)	76.72 (70.37–82.54)	68.42 (57.89–77.63)	82.30 (74.34–89.38)	72.22 (63.75–80.89)	79.49 (74.19–84.96)
Without I	187	80.92% (74.26%–86.91%)	77.54 (71.66–83.42)	67.11 (56.58–77.63)	84.68 (77.48–90.99)	75.00 (66.67–83.64)	78.99 (73.64–84.91)
Without C	189	54.98% (46.03%–64.04%)	56.61 (50.78–62.45)	23.68 (14.47–32.89)	78.76 (70.80–85.84)	42.86 (29.41–56.52)	60.54 (56.82–64.29)
Without T	189	76.58% (68.93%–83.14%)	73.02 (66.14–79.37)	60.53 (50.00–72.37)	81.42 (73.45–88.50)	68.66 (59.41–78.57)	75.41 (70.23–81.36)
I only	189	58.77% (50.41%–67.18%)	56.61 (49.74–64.02)	50.00 (39.47–60.56)	61.06 (52.21–69.91)	46.34 (38.16–54.76)	64.49 (58.33–71.13)
C only	187	77.76% (70.89%–84.20%)	68.98 (62.57–75.94)	63.16 (52.63–73.68)	72.97 (63.96–81.08)	61.54 (53.66–70.13)	74.31 (68.70–80.58)
T only	35	72.35% (54.92%–87.88%)	62.86 (48.57–77.14)	45.83 (25.00–66.67)	100.00 (100.00–100.00)	100.00 (100.00–100.00)	45.83 (37.93–57.89)

source biopsy narratives contain mixed findings across components, sides or time points; the strict-agreement extraction rule leaves HR and HER2 unknown. With the other clinical fields and treatment unavailable, the all-unknown clinical template is retained for provenance but excluded from model input. The image expert supplies their predictions.

Table 10: **Exact final-model modality restrictions.** Retained heads are intersected with each patient's observed-input eligibility. Every row uses the original frozen L3 weights.

Configuration	Removed heads	Retained heads
Without MRI/history	I, IC, IT, ICT	C, T, CT
Without clinical	C, IC, CT, ICT	I, T, IT
Without treatment	T, IT, CT, ICT	I, C, IC
MRI/history only	All except I	I
Clinical only	All except C	C
Treatment only	All except T	T

E Population allocation and fixed-order case diagnostics

Supplementary Figure S3 displays the complete-system and clinical–treatment comparisons. Supplementary Figure S4 summarizes frozen prediction dependence and native expert allocation. Supplementary Figure S6 presents the two fixed-order cases and their internal computations.

Population panels. Supplementary Figure S4(a) shows $\mathrm{AUC}_{\mathrm{full}} - \mathrm{AUC}_{\mathrm{restricted}}$ on matched eligible records; negative values favor the restriction. Hospital-B without-I uses 187 records and full AUC 80.07%; without-C/without-T use 189. Panel (b) averages native routing mass over baseline records, including zero mass for unavailable experts. Source records use their outer model; hospital records first average the five models' weights. Hospital-B allocation uses all 197 input records. The clinical-only expert receives approximately 0.53 mean weight in this normalized, availability-masked probability mixture.

Deterministic case selection and MRI panel. Each outcome stratum contributes its first eligible Hospital-A baseline record in fixed manifest order, with an observed mask and treatment units. The displayed axial slice maximizes mask area. The saved checkpoint, preprocessing and input tensors determine the computation. Grad-CAM (Selvaraju et al., 2017) targets the final pCR probability at

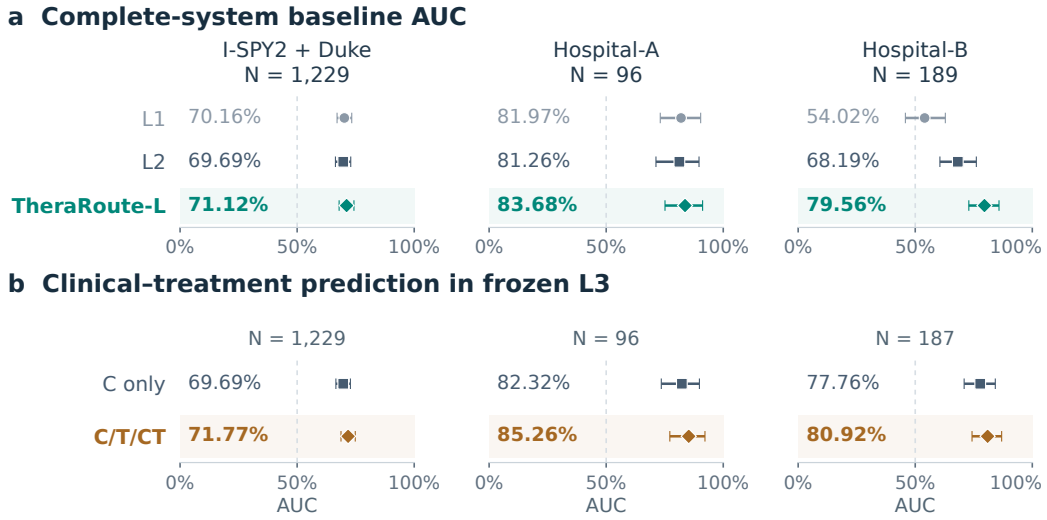


Figure S3: **Complete-system performance and clinical–treatment prediction.** (a) L-family baseline AUCs from Table 18. (b) Frozen C-only and C/T/CT predictions on matched populations. Hospital-B uses 189 records in (a) and 187 in (b). Whiskers reproduce the reported 95% conditional bootstrap intervals. All panels use the full AUC scale. Panel (a) compares complete systems; panel (b) restricts the same frozen L3 models. Appendix D reports the full mixture and all modality restrictions.

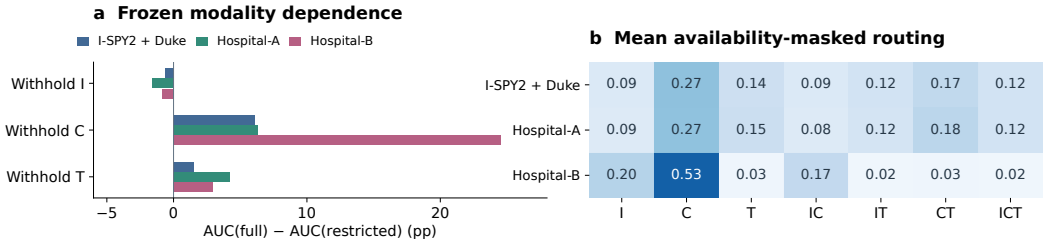


Figure S4: **Frozen prediction dependence and expert allocation.** (a) Full-minus-restricted AUC on matched eligible records; a negative difference favors the restriction. Hospital-B without MRI uses 187 records, whereas its other two restrictions use 189. (b) Mean availability-masked expert weights over baseline inputs, including zero weights for unavailable experts. Hospital-B routing averages all 197 inputs.

the third stage's signed spectral feature $A_c(v)$:

$$w_c = |\Omega|^{-1} \sum_{v \in \Omega} \frac{\partial p}{\partial A_c(v)}, \qquad H(v) = \max\{0, \sum_c w_c A_c(v)\}. \tag{59}$$

Each map is trilinearly upsampled to the input grid. Five raw maps are averaged, then normalized per case by the three-dimensional maximum. Both cases use the positive pCR target. Grad-CAM describes local probability sensitivity at the selected stage; spatially broad responses extend beyond the observed lesion contour. The panel displays these derivatives alongside the separate observed mask. CPU re-evaluation of ten frozen case/model pairs reproduces saved probabilities within 2.09×10^{-6}, with parameters unchanged.

Treatment and probability panels. Supplementary Figure S6(b) displays the unit-read weights $\eta^U_{ie,j}$ defined in Equation (3), separately for IT, CT and ICT and averaged across five models. Both cases share docetaxel, doxorubicin and cyclophosphamide units and three concurrent relations. Their different reads illustrate patient conditioning with a fixed unit set. The display focuses on the query pathway; whole-description and structured features additionally enter the expert classifier and raw

routing head. Clinical slot presence is uniform over observed slots, so the figure displays the learned treatment reads and keeps raw clinical/pathology text restricted.

Panel (c) pairs the clinical expert and full-mixture probabilities for each model and their equal-weight means: 0.1998 to 0.2713 for Case 1 (non-pCR), and 0.4225 to 0.3498 for Case 2 (pCR). Both updates move away from the observed endpoint; Case 2 remains a false negative at the Hospital-A operating point. These fixed-order cases expose the internal updates and their errors alongside the cohort-level performance tables.

Hospital-B protocol. Ensembling, local threshold fitting and endpoint eligibility follow Section 3.2.

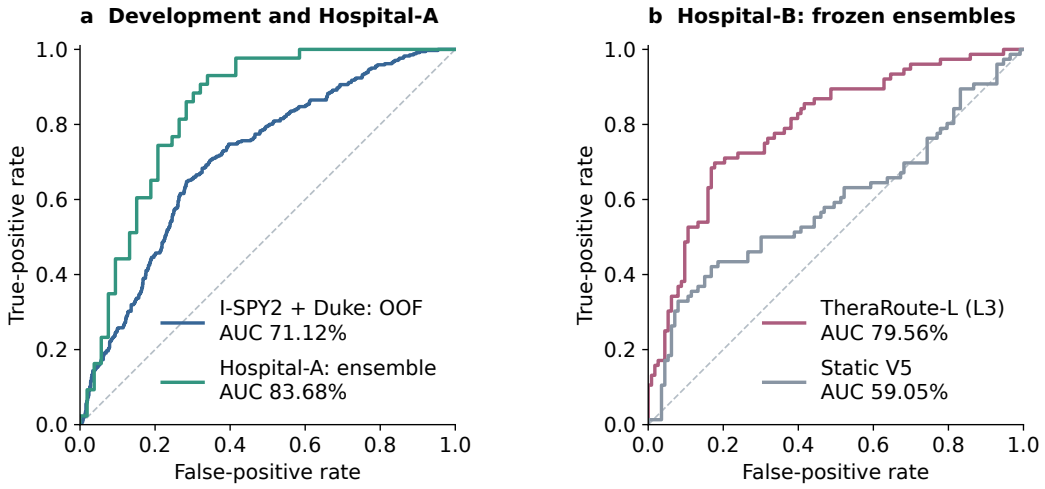


Figure S5: **Empirical baseline ROC curves.** (a) Combined development OOF probabilities and the Hospital-A frozen ensemble. (b) Final TheraRoute-L and static V5 five-model equal-weight ensembles in Hospital-B, evaluated with the five-fold cross-fitted Youden procedure for classification. Curves are unsmoothed and use the stored continuous scores.

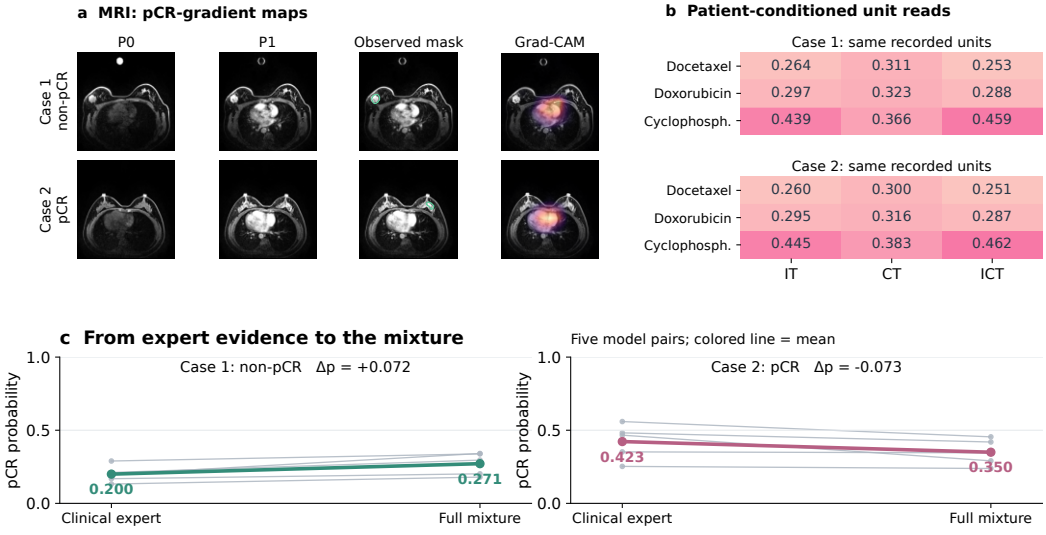


Figure S6: **Fixed-order case diagnostics.** (a) P0, P1, observed masks and third-stage pCR Grad-CAM averaged across five frozen models. (b) IT, CT and ICT unit-query weights for a shared treatment-unit set. (c) Clinical-expert and full-mixture probabilities, with equal-weight means. Each outcome stratum contributes its first eligible Hospital-A record in fixed manifest order. Appendix E reports the probability updates and outcomes.

F CONTROLLED ABLATIONS AND HISTORICAL COMPARISONS

F.1 CONTROLLED TREATMENT-REPRESENTATION ABLATIONS

Three retrained variants. A0 is the complete TheraRoute-L model. A1 retains whole-regimen text and removes explicit unit and relation pooling together with the dedicated treatment reads. A2 retains whole-regimen text, unit pooling and patient-conditioned unit reads. It removes relation pooling, static relation reads and interval relation reads. The original descriptions, including words expressing combination or order, are unchanged. All variants retain MRI, clinical fields, history encoding, the seven experts and routing.

Fixed training conditions. Each variant is trained from initialization on the same five source splits, giving 15 completed runs. Shared parameters start from the same values. All model modules are constructed in the same order; bypassed ablation parameters are frozen and excluded from the optimizer. The input manifest, labels, training visits, patient weights, augmentation, optimizer, loss coefficients and training schedule are fixed. The runs use FP32, physical batch size two, eight accumulation steps, six workers, maximum 100 epochs and patience 15. Split seed is 42 and initialization seed is $42 + k$ for fold k. Each variant selects its checkpoint with the same source-inner rule. No loss weight is reassigned when a pathway is removed: A1's dedicated response term and its penalty are zero, while A2 retains the unit-response term with its original penalty coefficient.

Paired evaluation. Development uses one outer-held-out score per baseline record and the corresponding source-inner cohort-by-task cutoff. Hospital-A uses equal-weight probabilities from five frozen models and cutoff 0.5. Hospital-B uses the same 189 resolved endpoints and local five-fold assignments for every variant. Each variant fits its own Youden cutoffs on the other four local folds. AUC uses continuous probabilities. Tables 12–14 give all six metrics; Table 15 gives all paired AUC differences.

A0–A1 compares the full explicit treatment pathway with whole-regimen semantics; A0–A2 removes relation pooling and reads; A2–A1 adds unit pooling and reads. Query design, relation rank and longitudinal training are held fixed. All active pathways are jointly retrained, so these contrasts describe incremental configurations within the shared system. The small source differences and population-dependent hospital changes are consistent with complementary representations, without establishing large independent module benefits or a statistical interaction effect. These comparisons complement the frozen expert restrictions in main-text Table 2. In the tables, W denotes whole-regimen text, U explicit units and their reads, and R explicit relations with pooling and reads. Checkmarks denote active components; bold marks the highest displayed point estimate in each metric column, including ties.

Table 11: **Controlled ablations of treatment representation.** Five source-fold runs per variant; units and relations include pooling and reads. All retain MRI, clinical inputs, history and experts. Parentheses: 95% conditional bootstrap intervals; bold: highest AUC. Source scores are outer-held-out; hospital scores are five-model means.

	Treatment components			AUC (95% CI)		
Variant	Whole text	Units	Relations	I-SPY2 + Duke $N = 1{,}229$	Hospital-A $N = 96$	Hospital-B $N = 189$
A1	✓	–	–	71.31% (68.15%–74.36%)	82.40% (73.28%–89.95%)	76.01% (68.55%–83.09%)
A2	✓	✓	–	**71.37%** (68.20%–74.51%)	82.93% (73.98%–90.74%)	77.99% (71.07%–84.83%)
A0	✓	✓	✓	71.12% (67.93%–74.26%)	**83.68%** (75.12%–91.18%)	**79.56%** (72.89%–85.78%)

Fixed-threshold results. Hospital-A A1 has $TP = 0$, $FN = 43$, $TN = 52$ and $FP = 1$. Its sensitivity and PPV are zero. A2 and A0 each have $TP = 34$, $FN = 9$, $TN = 39$ and $FP = 14$, giving identical threshold-based metrics despite different AUCs. The zero-width A1 sensitivity and PPV intervals follow the fixed-decision bootstrap; PPV is omitted from bootstrap draws with no predicted positive. The complete results retain these operating-point differences.

Table 12: **Controlled treatment ablations: I-SPY2 + Duke.** The same 1,229 baseline records are evaluated once by their outer-held-out models. Classification uses source-inner cohort-by-task Youden cutoffs. AUC is a percentage; other metrics are percentages. Parentheses give 95% conditional bootstrap intervals.

	Components			Performance (95% CI)					
Variant	W	U	R	AUC	ACC	SEN	SPE	PPV	NPV
A1	✓	–	–	71.31% (68.15%–74.36%)	**65.66** (63.06–68.35)	68.33 (63.05–73.02)	**64.64** (61.48–67.57)	42.60 (39.89–45.52)	84.16 (81.94–86.41)
A2	✓	✓	–	**71.37%** (68.20%–74.51%)	65.26 (62.65–68.02)	**71.55** (66.86–75.96)	62.84 (59.35–66.22)	42.51 (40.03–45.34)	**85.19** (83.08–87.38)
A0	✓	✓	✓	71.12% (67.93%–74.26%)	65.58 (62.89–68.35)	70.09 (65.40–74.78)	63.85 (60.59–67.12)	**42.68** (40.03–45.65)	84.75 (82.57–86.92)

Table 13: **Controlled treatment ablations: Hospital-A.** Equal-weight five-model probabilities at the fixed 0.5 cutoff, $N = 96$. AUC is a percentage; other metrics are percentages. Parentheses give 95% conditional bootstrap intervals.

	Components			Performance (95% CI)					
Variant	W	U	R	AUC	ACC	SEN	SPE	PPV	NPV
A1	✓	–	–	82.40% (73.28%–89.95%)	54.17 (52.08–55.21)	0.00 (0.00–0.00)	**98.11** (94.34–100.00)	0.00 (0.00–0.00)	54.74 (53.76–55.21)
A2	✓	✓	–	82.93% (73.98%–90.74%)	**76.04** (67.71–84.38)	**79.07** (67.44–90.70)	73.58 (62.22–84.91)	**70.83** (61.11–81.40)	**81.25** (72.92–90.91)
A0	✓	✓	✓	**83.68%** (75.12%–91.18%)	**76.04** (67.71–84.38)	**79.07** (67.44–90.70)	73.58 (62.22–84.91)	**70.83** (61.11–81.40)	**81.25** (72.92–90.91)

Table 14: **Controlled treatment ablations: Hospital-B.** Equal-weight five-model probabilities and five-fold cross-fitted Youden cutoffs, using the same 189 pathology endpoints. AUC is a percentage; other metrics are percentages. Parentheses give 95% conditional bootstrap intervals.

	Components			Performance (95% CI)					
Variant	W	U	R	AUC	ACC	SEN	SPE	PPV	NPV
A1	✓	–	–	76.01% (68.55%–83.09%)	73.02 (66.67–78.85)	**69.74** (59.21–78.98)	75.22 (66.37–82.30)	65.43 (57.78–73.53)	78.70 (72.65–84.91)
A2	✓	✓	–	77.99% (71.07%–84.83%)	69.31 (62.43–75.66)	65.79 (55.26–76.32)	71.68 (61.95–80.53)	60.98 (53.09–69.62)	75.70 (69.79–81.82)
A0	✓	✓	✓	**79.56%** (72.89%–85.78%)	**76.72** (70.37–82.54)	68.42 (57.89–77.63)	**82.30** (74.34–89.38)	**72.22** (63.75–80.89)	**79.49** (74.19–84.96)

For Hospital-B, A0–A1 and A0–A2 conditional intervals lie above zero; A2–A1 includes zero. All development and Hospital-A paired intervals include zero. These population-specific increments are interpreted together with the coupled learning paths in Section 2.5 and the joint-representation discussion in Section 4.3.

F.2 Retrained published-model comparators

Methods and provenance. We retrained the authors' released implementations of two published response-prediction architectures. MRP, the Multi-modal Response Prediction system of Gao

Table 15: **Paired AUC differences between jointly retrained treatment configurations.** First variant minus second, using identical records and 1,000 paired class-stratified bootstrap resamples. Intervals condition on frozen predictions and are unadjusted for multiple comparisons.

Group	Comparison	ΔAUC (pp)	95% CI (pp)
I-SPY2 + Duke	A0–A1	-0.19	$(-0.95, +0.64)$
	A0–A2	-0.25	$(-0.68, +0.20)$
	A2–A1	$+0.06$	$(-0.75, +0.95)$
Hospital-A	A0–A1	$+1.27$	$(-0.92, +3.73)$
	A0–A2	$+0.75$	$(-0.53, +2.06)$
	A2–A1	$+0.53$	$(-1.80, +2.85)$
Hospital-B	A0–A1	$+3.55$	$(+0.12, +7.15)$
	A0–A2	$+1.57$	$(+0.34, +2.89)$
	A2–A1	$+1.98$	$(-2.07, +6.00)$

et al. (2024) in *Nature Communications*, combines mammography ($iMGrhpc$) and MRI–clinical ($iMRrhpc$) branches; ***MRP–MRI*** denotes the latter branch evaluated here. MORM is the authors' repository name for the multimodal probabilistic response model of Han et al. (2025) in *eClinicalMedicine*. Table 1 reports both implementations retrained and evaluated on the study populations.

Shared data and evaluation. Both comparators use the same source-patient folds, source-inner partitions, reviewed endpoints and MRI preprocessing as TheraRoute-L. MRI inputs have size $88 \times 176 \times 176$. The primary analysis uses baseline observations only. Source results pool one outer-held-out prediction per patient for I-SPY2 + Duke ($N = 1{,}229$); classification uses the source-inner cohort-specific Youden thresholds. Hospital-A ($N = 96$) uses the equal-weight probability mean of five source models and a threshold of 0.5. Hospital-B uses the same five-model probability mean and the locked, separate five-fold cross-fitted Youden procedure on 189 resolved endpoints; eight of its 197 input records remain outside endpoint scoring. The six metrics and their existing conditional 95% intervals use the manuscript's fixed-prediction and fixed-decision resampling procedure. All model-selection steps use source data.

MRP–MRI input and supervision. The shared 3D ResNet-18, two MRI slots, temporal-state embeddings, 27 scalar clinical/treatment slots, main pCR head, image-only pCR head and observed-attribute heads are retained. Each examination supplies its signed P1–P0 enhancement volume. The two slots contain the first and current examination of an observed prefix; at baseline, the same examination occupies both slots, with baseline time codes. Available source histories contribute patient-normalized prefix losses, while source-inner model selection and reported metrics use baseline predictions. Common fields include age, HR, HER2, subtype, menopause and explicitly recorded ER/PR. Two scalar slots originally assigned to unavailable cN/cM fields are remapped to HR/subtype; unknown scalar inputs are encoded as -1. The observed auxiliary targets are age, menopause, recorded treatment categories and subtype. The main and image-only binary losses each have weight 10; auxiliary weights are 0.0001 for age, 0.01 for menopause, 0.025 for each recorded treatment category and 1 for subtype. Auxiliary losses use observed targets.

MORM input and representation. The retained architecture comprises a 33-dimensional tabular input, mask-guided 3D ResNet-10 with attention pooling, three text slots, 256-dimensional hidden features, three-dimensional Gaussian latent variables, precision-weighted Product-of-Experts fusion and the regimen-conditioned classifier with its correction gate. The MRI branch receives P0/P1 channels and a separate mask; absent masks are zero. The original RadioLOGIC/RadioBERT text backbone is replaced by the same local BioBERT specification used for TheraRoute-L clinical encoding. BioBERT stays frozen and in evaluation mode, and its CLS features enter the original text projection. Available clinical descriptions fill the patient-record slot; unavailable radiology and diagnostic-pathology report slots are zero. Records without clinical text use the tabular–image fusion path. The pCR-defining postoperative pathology is reserved for endpoint construction. The 33 tabular entries contain 17 typed value positions, seven observation indicators and nine reserved zeros.

MORM treatment mapping and objective. The original five treatment categories are anthracycline backbone, anthracycline–taxane backbone, single HER2 antibody, multiple HER2 antibodies and hormone therapy. For the available cross-cohort records, the five-dimensional interface instead

encodes recorded cytotoxic, endocrine, HER2-targeted, checkpoint-inhibitor and other-targeted treatment categories. An entirely missing or unmapped regimen is encoded as -1; zero for an individual category means that it was not recorded. The pCR-only retraining objective combines the available pCR heads and Gaussian KL terms, replacing the original joint pCR/survival objective. Both comparators use randomly initialized MRI encoders. MRP–MRI encodes its clinical/treatment inputs as scalars; MORM uses the frozen language backbone specified above.

Optimization. MRP–MRI uses Adam with learning rate 10^{-4}, weight decay 10^{-6}, gradient clipping at 0.25, physical batch size 2 and four-step accumulation. Its maximum budget is 100 epochs with early-stopping patience 20; a validation-loss plateau scheduler is updated every three epochs. MORM first trains modality alignment separately within each source-training fold for 100 epochs at batch size 4, retaining six directed contrastive terms and the observed single-modality KL terms. The lowest source-training alignment loss selects its encoder checkpoint. Frozen features then feed the PoE phase, with batch size 655, a 500-epoch maximum and patience 50. Both MORM phases use AdamW with learning rate 10^{-4}, weight decay 0.05 and betas $(0.9, 0.95)$; class-balanced sampling uses source-training labels. Final pCR checkpoints are selected by pooled source-inner baseline AUC, with binary cross-entropy as the tie-breaker. Inference uses latent means. The comparison therefore shares analysis populations and scoring rules while retaining architecture-specific inputs, training objectives and schedules.

Undefined predictive value. MRP–MRI's all-negative Hospital-A decisions at 0.5 give an undefined PPV and interval, shown as N/A. Class-stratified resampling of these fixed decisions gives zero-width ACC, SEN, SPE and NPV intervals, as reported in Table 1.

F.3 Relationship to recent literature

Han et al. (2025) combine pre-NAT clinical data, DCE-MRI and medical reports to predict pCR and survival under observed regimens. MLRL uses vision–text alignment within examinations and visual/textual progression objectives across examinations (Gao et al., 2025). The MRP study of Gao et al. (2024) combines longitudinal multimodal observations with time embeddings and clinical attributes, while Ma et al. (2025a) integrate subtype-aware attention and treatment-associated longitudinal MRI changes. Tumor-preserving registration provides a complementary approach to quantifying longitudinal DCE-MRI changes (Han et al., 2026). TheraRoute-L focuses on shared treatment units, explicit relation types and patient-conditioned reads, with distinct within-examination and between-visit pathways.

Compositional structure also appears in perturbation modeling: NAIAD combines individual perturbation effects with adaptive embeddings to prioritize combinatorial experiments (Qin et al., 2025). Its experimental effect-prediction target differs from the clinical pCR endpoint here. In clinical representation learning, MUSE uses a patient–modality graph and mutual-consistency objectives to accommodate missing modalities and labels (Wu et al., 2024). In TheraRoute-L, missing modalities are expressed through explicit expert eligibility.

MOME adapts modality experts to multiparametric breast MRI (Luo et al., 2025), while M4oE models patient-dependent modality/task relationships (Wu et al., 2025). MRICombo uses multiple experts for heterogeneous MRI sequences and segmentation and tumor-characterization tasks (Zhang et al., 2026). Our learned routing weights allocate probability mass among available private modality-subset experts. R2-T2 examines test-time expert rerouting in large multimodal models (Li et al., 2025); TheraRoute-L trains its router through direct supervision and source training-risk targets, then freezes it for hospital prediction. Hospital-B adapts the classification cutoff through local cross-fitting.

Alternating unimodal adaptation uses a shared head and gradient modification to address optimization interference (Zhang et al., 2024); Data Remixing works through training-data reassembly (Ma et al., 2025b). Asymmetric Representation Learning further shows why equal modality influence need not be optimal (Wei et al., 2025). TheraRoute-L uses direct subset-expert supervision, patient weighting and staged routing. For absent inputs, duplex diffusion provides a generative alternative (Dai et al., 2025); our implementation restricts expert availability to observed inputs. Frequency-modulated vision encoders also motivate explicit spectral selectivity (Lin et al., 2025), with the present implementation specialized to signed 3D phase differences.

The following tables compare historical systems under their documented training and selection protocols.

F.4 SEVEN COMPLETE-SYSTEM DESIGNS

Table 16: **Historical design context.** Rows compare complete systems with multiple design changes. Static V2 and longitudinal L2 denote distinct architectures.

System	Implemented design and protocol distinction
V1	Early modality interaction before expert heads; channel-wise frequency weights. Its source held-out set participated in epoch selection.
V2	Radial MRI frequency bands, no text FFT, seven private modality-subset experts with stochastic latents, stable frozen text encoding. Source epoch selection used the reported held-out set.
V4	Shared phases, signed differences and constrained image increments; staged expert/router fitting and source calibration. Its training and probability protocol differ from L3.
V5	Joint training, a pre-fixed EMA endpoint and source-only monotonic calibration.
L1	Deterministic private modality-subset experts, compositional treatment and history support; clinical typed-field use and route balance were limited.
L2	Explicit typed fields, missing-input augmentation and a baseline-preserving temporal path; a fixed image–clinical fusion share remains.
L3	Shared-phase energy modulation, current-visit identity, typed clinical context, unit/relationship treatment encoding and learned available-expert fusion; source-inner selection and cohort-by-task thresholds.

Static V4 fits experts, calibration and the router in stages with a 99-record source calibration partition. Static V5 uses joint fusion training, a pre-fixed 100-epoch EMA endpoint and source-fitted positive-slope probability calibration. Evaluation uses these final models. All seven systems share Hospital-B post-processing while preserving their frozen preprocessing and source calibration.

Table 17: **Historical L-family complete systems: development and Hospital-A.** All six original metrics are retained. Development uses pooled outer predictions and source-inner cutoffs, task-specific for L1/L2 and cohort-by-task for L3. Hospital-A uses five-model probability means at 0.5. AUC is a percentage; the other metrics are percentages. L1's Hospital-A PPV is undefined because no record is predicted positive.

System	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
I-SPY2 + Duke ($N = 1229$)						
L1	70.16% (67.05%–73.32%)	65.17 (62.57–67.94)	67.74 (63.05–72.73)	64.19 (61.26–67.45)	42.08 (39.44–45.02)	83.82 (81.83–86.12)
L2	69.69% (66.36%–72.89%)	66.97 (64.04–69.65)	56.60 (51.32–61.88)	70.95 (67.79–73.87)	42.79 (39.24–46.35)	80.98 (78.97–83.09)
TheraRoute-L (L3)	71.12% (67.93%–74.26%)	65.58 (62.89–68.35)	70.09 (65.40–74.78)	63.85 (60.59–67.12)	42.68 (40.03–45.65)	84.75 (82.57–86.92)
Hospital-A ($N = 96$)						
L1	81.97% (73.10%–90.35%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
L2	81.26% (71.26%–89.65%)	55.21 (48.96–61.46)	11.63 (2.33–20.93)	90.57 (81.13–98.11)	50.00 (19.95–81.86)	55.81 (52.44–59.26)
TheraRoute-L (L3)	83.68% (75.12%–91.18%)	76.04 (67.71–84.38)	79.07 (67.44–90.70)	73.58 (62.22–84.91)	70.83 (61.11–81.40)	81.25 (72.92–90.91)

Table 18: **AUC comparison of L-family complete systems.** Source results use pooled outer-fold predictions. Hospital results use equal-weight probabilities from five source models. Parentheses give 95% conditional bootstrap intervals. Table 16 details the system configurations; Table 3 in the main text reports all seven Hospital-B systems and all six metrics.

Complete system (design summary)	I-SPY2 + Duke 95% CI	Hospital-A 95% CI	Hospital-B 95% CI
L1: earlier deterministic subset experts	70.16% (67.05%–73.32%)	81.97% (73.10%–90.35%)	54.02% (45.80%–62.86%)
L2: typed fields, fixed image–clinical share	69.69% (66.36%–72.89%)	81.26% (71.26%–89.65%)	68.19% (60.54%–76.10%)
TheraRoute-L (L3): learned available-expert fusion	71.12% (67.93%–74.26%)	83.68% (75.12%–91.18%)	79.56% (72.89%–85.78%)

Table 19: **Historical static complete systems: combined development.** V1/V2 use the reported held-out source fold for epoch and threshold selection. V4 uses inner selection/refitting/calibration; V5 uses fixed-epoch EMA and inner calibration. Table 17 specifies the L-family protocols.

Configuration	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
V1	71.01% (67.83%–74.33%)	70.46 (67.94–72.82)	68.04 (62.76–73.02)	71.40 (68.24–74.32)	47.74 (44.63–50.76)	85.33 (83.33–87.29)
V2	69.66% (66.45%–72.88%)	69.08 (66.48–71.60)	72.73 (67.45–77.42)	67.68 (64.53–70.95)	46.36 (43.58–49.25)	86.60 (84.41–88.64)
V4	66.80% (63.36%–70.09%)	59.89 (57.28–62.49)	66.28 (61.58–71.55)	57.43 (54.05–60.36)	37.42 (35.09–39.93)	81.60 (79.37–84.10)
V5	61.49% (57.95%–64.85%)	54.35 (51.58–57.12)	69.21 (64.22–74.49)	48.65 (45.38–51.80)	34.10 (31.91–36.36)	80.45 (77.56–83.30)

Table 20: **Historical static complete systems: Hospital-A.** Five-model mean probabilities at the common 0.5 cutoff, using the fixed 96-record order.

Configuration	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
V1	62.18% (51.59%–74.03%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
V2	78.28% (68.45%–87.01%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
V4	66.30% (55.15%–76.92%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
V5	57.17% (45.76%–68.89%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)

F.5 Five independently trained modality/fusion baselines

The V1 experimental suite contains five independently trained configurations. Image-only supplies paired MRI and a tumor mask to the volumetric encoder and a classifier. Clinical-only uses the clinical prompt encoder and its original text FFT. Treatment-only uses the treatment prompt encoder and its original text FFT. These single-modality configurations receive their named modality alone. Three-modal concatenation uses the original image/text branches and an MLP; those text branches retain their original clinical–treatment interaction. Three-modal MHA lets image tokens attend to clinical/treatment tokens and classifies the pooled image and text features. Tables 21 and 22 report these independently trained configurations.

Each configuration has five seed-42 runs with best checkpoints and source/Hospital-A predictions. The reported source fold also selects the best epoch by AUC; L3 instead uses a separate source-inner selection partition. Hospital-A predictions follow the ordered 96-record manifest and are averaged across models; source predictions are pooled once per held-out fold.

Table 21: **Independently trained historical baselines: combined development.** Five runs per configuration. Source cutoffs and epoch selection use the reported held-out folds. AUC is pooled from probabilities; the other five metrics use pooled decisions.

Configuration	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
Image only	56.71% (53.35%–60.15%)	57.36 (54.84–60.13)	74.49 (69.79–78.89)	50.79 (47.52–53.94)	36.76 (34.81–38.93)	83.83 (81.41–86.36)
Clinical only	64.83% (61.39%–68.12%)	60.21 (57.61–62.98)	73.31 (68.33–78.01)	55.18 (52.03–58.45)	38.58 (36.38–40.92)	84.34 (81.89–86.75)
Treatment only	69.32% (66.09%–72.51%)	65.83 (63.14–68.51)	63.05 (57.77–68.33)	66.89 (63.74–69.93)	42.24 (39.31–45.39)	82.50 (80.41–84.79)
Three-modal concat	72.81% (69.46%–75.87%)	70.79 (68.10–73.23)	65.69 (60.70–70.67)	72.75 (69.82–75.45)	48.07 (44.70–51.37)	84.67 (82.74–86.64)
Three-modal MHA	70.46% (67.14%–73.66%)	68.84 (66.31–71.44)	71.85 (67.16–76.25)	67.68 (64.64–70.72)	46.05 (43.18–49.01)	86.23 (84.17–88.32)

Table 22: **Independently trained historical baselines: Hospital-A.** Equal-weight five-model probabilities, fixed 0.5 cutoff, $N = 96$. N/A denotes an undefined predictive value with zero predicted positives.

Configuration	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
Image only	50.20% (37.42%–61.74%)	54.17 (50.00–58.33)	4.65 (0.00–11.63)	94.34 (86.79–100.00)	40.00 (0.00–100.00)	54.95 (52.75–57.30)
Clinical only	50.68% (39.00%–62.40%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Treatment only	67.07% (55.76%–77.71%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Three-modal concat	67.62% (56.56%–78.06%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Three-modal MHA	70.95% (60.94%–81.44%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	N/A	55.21 (55.21–55.21)

F.6 Eight historically retrained module variants

The V1 removal/replacement suite contributes another 40 completed runs. Removing image FFT removes spectral attention and global/window Fourier blocks while retaining convolution and mask attention. Removing mask attention retains the spectral path but removes lesion boosting. Removing text FFT substitutes padding-masked identity. Removing text interaction replaces the bidirectional clinical–treatment FiLM maps with masked identity. Removing frequency cross-modal fusion preserves and pools image tokens without text conditioning. Replacing PoMoE retains the preceding encoders/fusion and substitutes a concatenation classifier for latent expert fusion. Its objective consequently omits unavailable expert-head terms. Removing multi-branch supervision preserves the architecture but removes unimodal/bimodal CE. Removing FFT regularization changes its coefficient to zero while retaining the architecture and branch supervision.

Tables 23 and 24 report all eight V1 interventions under their original source-selection and threshold conventions.

Table 23: **Historically retrained V1 module variants: combined development.** Five runs per row; the reported source folds also select the epoch and threshold. The V1 reference appears in Table 19.

Configuration	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
Remove image FFT	67.30% (64.05%–70.74%)	66.88 (64.28–69.49)	74.19 (69.79–78.89)	64.08 (60.81–67.12)	44.23 (41.62–46.90)	86.61 (84.57–88.87)
Remove mask attention	71.66% (68.44%–74.87%)	67.94 (65.25–70.47)	71.85 (67.16–76.25)	66.44 (63.18–69.37)	45.12 (42.31–47.97)	86.01 (83.88–88.02)
Remove text FFT	71.53% (68.54%–74.69%)	66.31 (63.79–68.76)	75.07 (70.67–79.77)	62.95 (59.68–65.99)	43.76 (41.33–46.18)	86.80 (84.74–89.10)
Remove text interaction	70.46% (67.11%–73.79%)	67.86 (65.26–70.46)	72.14 (67.45–76.83)	66.22 (63.18–69.15)	45.05 (42.25–47.98)	86.09 (84.08–88.32)
Remove FFT cross-modal fusion	70.85% (67.71%–74.19%)	65.50 (63.06–68.02)	76.25 (71.85–80.65)	61.37 (58.45–64.53)	43.12 (40.82–45.53)	87.06 (85.05–89.20)
Replace PoMoE by concat	72.50% (69.50%–75.73%)	69.24 (66.80–71.85)	71.55 (66.57–76.54)	68.36 (65.32–71.28)	46.48 (43.66–49.52)	86.22 (84.13–88.43)
Remove branch supervision	69.74% (66.36%–73.38%)	68.43 (65.74–70.95)	71.85 (67.16–76.54)	67.12 (64.08–70.16)	45.62 (42.64–48.51)	86.13 (84.02–88.25)
Remove FFT regularizer	72.19% (68.87%–75.42%)	69.16 (66.56–71.60)	71.55 (66.57–76.54)	68.24 (65.20–71.06)	46.39 (43.54–49.24)	86.20 (84.11–88.25)

Table 24: **Historically retrained V1 module variants: Hospital-A.** Five-model mean probabilities, common 0.5 cutoff, fixed $N = 96$.

Configuration	AUC 95% CI	ACC 95% CI	SEN 95% CI	SPE 95% CI	PPV 95% CI	NPV 95% CI
Remove image FFT	62.79% (51.34%–73.58%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Remove mask attention	68.80% (57.70%–79.25%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Remove text FFT	72.62% (62.31%–82.49%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Remove text interaction	68.63% (57.92%–79.20%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Remove FFT cross-modal fusion	74.55% (64.15%–84.38%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Replace PoMoE by concat	70.51% (60.24%–81.05%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Remove branch supervision	71.22% (60.64%–81.36%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)
Remove FFT regularizer	70.95% (60.24%–81.62%)	55.21 (55.21–55.21)	0.00 (0.00–0.00)	100.00 (100.00–100.00)	*N/A*	55.21 (55.21–55.21)

F.7 Historical run inventory

The historical archive comprises 13 configurations: five modality/fusion baselines and eight V1 module variants, each trained across five folds, for 65 runs in total. Complete-system designs are summarized separately in Table 16.

G Study provenance and interpretation

Endpoint sources. Reference annotations use cohort-specific source fields. I-SPY2 uses the source clinical table's binary pCR field; Hospital-A uses the institutional worksheet's binary pCR field in the fixed 96-record order. Duke uses post-neoadjuvant pathologic T/N fields from the initial surgical intervention, excluding re-excisions according to its source dictionary. Hospital-B uses the pathology-text rule in Appendix B.1; post-treatment pathology supplies the endpoint only.

Duke endpoint coding. The retained Duke loader applies an implementation-level binary conversion to the parseable T/N fields: T0/Tis with N0 is positive, and all other combinations are negative. Numerical -1 denotes TX or NX and is parseable.

Physician review confirmed the existing non-pCR reference labels for the 28 Duke records identified during source-field checks, including the 24 records with a complete-response entry in a separate source field. All 28 reference labels remained unchanged. Original source fields are preserved for traceability.

Source-to-inclusion traceability. Final manifest records retain original source fields and ordered links to source identifiers and historical reference-label export rows. The *Implementation and Scoring Guide* identifies the frozen files by SHA-256 fingerprints and documents the scoring populations; the anonymous *Endpoint Evidence* attachment (`reproducibility/endpoint_evidence.json`) reports the endpoint coding and the 28-record clinical confirmation, with patient-level links kept under controlled access. Hospital development and threshold adaptation follow Section 3.2.

Institutional documentation. Institutional approvals were obtained, and approval identifiers and informed-consent or waiver documentation were finalized before submission. Identifying institutional details are withheld from this anonymous manuscript.