JOINT MRI ACQUISITION PATTERNS CHANGE MODEL SEPARATION IN BRAIN-TUMOR SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain-tumor MRI sequences arrive in workflow bundles: availability of T1, T1Gd, T2, and FLAIR is jointly structured, whereas standard incomplete-modality tests often hide each sequence independently. We introduce observability conclusion sensitivity (OCS), the change in a fixed model pair's performance margin between two mask distributions, to measure whether the evidence separating methods survives that benchmark choice. Our evaluation freezes a joint prior estimated from 9,842 Centers A/B workflow logs, shares patient and mask draws within comparisons, separates training from evaluation exposure, and instantiates the test with ProtocolProbe, a Swin UNETR probe combining acquisition-state tokens, protocol-conditioned modality masking, and subset consistency. Against mmFormer, ProtocolProbe's margin increases from 2.4 Dice under Rand(0.5) to 4.3 under Protocol, giving OCS = 1.9 (paired 95% CI [0.9, 3.0], $p = 0.001$). After mmFormer receives the same protocol exposure, OCS remains 1.0 Dice (95% CI [0.3, 1.7]); a fixed-backbone ID-only control yields OCS = 0.1. The response spans 0.7–1.3 Dice across eight separately frozen prior constructions, while external performance margins over protocol-aware mmFormer are 2.4 Dice at both development-isolated Centers C and D. Thus, the full joint mask distribution is a claim-bearing part of an incomplete-modality benchmark: freezing it across methods and reporting its paired margin response reveals comparative evidence that a single i.i.d. dropout score conceals.

1 INTRODUCTION

Four-sequence brain-tumor segmentation is evaluated over sixteen possible acquisition states, yet the distribution over those states is usually treated as a simulator setting rather than part of the scientific comparison. This abstraction is consequential because T1, T1Gd, T2, and FLAIR arrive as protocol bundles shaped by contrast eligibility, abbreviated exams, site workflow, and image quality. When two fusion models fail differently on the same bundles, their average-score margin is a weighted quantity: changing the acquisition weights can change the evidence for preferring one model without changing either fitted model.

We make that weighting explicit by indexing every comparison with its observability distribution. For fixed fitted models a and b, *observability conclusion sensitivity* (OCS) is the difference between their margin under a prespecified workflow prior and their margin under Rand(0.5). The prior is estimated once from 9,842 de-identified Centers A/B modality-presence logs, frozen across model families, and applied with paired patients and shared within-regime missingness realizations. Separate indices for training and evaluation exposure distinguish the response of an existing comparison from gains due to protocol-aware training. This shifts the object of robustness analysis from one model's score to the comparative evidence carried by a benchmark.

Existing incomplete-modality methods improve individual predictors through dropout, reconstruction, distillation, or specialized fusion (Neverova et al., 2016; Havaei et al., 2016; Dorent et al., 2019; Zhang et al., 2022; Wang et al., 2023; Zeng et al., 2024; Kang et al., 2026; Zhu et al., 2025; Novosad et al., 2024). Missing-data research likewise studies prediction under changing observation mechanisms (Rubin, 1976; Little & Rubin, 2002; Zhou et al., 2023; Zhu et al., 2023; Zhu & Jiang, 2024; Rockenschaub et al., 2024). These lines establish how to build or adapt a predictor; OCS instead tests whether a method-to-method claim transports between two named joint mask distributions.

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.

Unlike a scalar dropout sweep, the intervention freezes the complete state distribution across methods, and unlike a rank-only check, the test retains the effect size and paired uncertainty used to interpret competitive separation.

ProtocolProbe instantiates this comparison with a shared Swin UNETR backbone, explicit ACQ/MISS/LOWQ state tokens, protocol-conditioned modality masking (PCMM), and subset-consistency regularization (SCR). The evaluation then attacks distinct alternative explanations: protocol-aware mmFormer and ShaSpec receive the same prior and masked training exposure; an ID-only variant preserves the backbone while removing the observability-coupled bundle; eight center, temporal, smoothed, and perturbed priors test prior construction; and Centers C and D provide development-isolated external performance validation. All comparisons retain common splits, optimization budgets, ensembles, post-processing, patients, and within-regime mask draws.

The resulting evidence forms a single conclusion-robustness profile. Relative to protocol-aware mmFormer, OCS is 1.0 Dice (paired 95% CI [0.3, 1.7]) and spans 0.7–1.3 across the eight prior constructions. The same pair is separated by 2.4 Dice at each development-isolated center, while the shared-backbone ID-only control produces OCS = 0.1 rather than ProtocolProbe's 1.9 against mmFormer. Together, the estimand, frozen cross-family intervention, and controlled MRI study establish that joint acquisition patterns determine how much comparative evidence an incomplete-modality benchmark reports.

2 Related Work

Incomplete-modality prediction. Modality dropout (Neverova et al., 2016), feature-statistics aggregation in HeMIS (Havaei et al., 2016), sequence completion in U-HVED (Dorent et al., 2019), and fusion architectures such as mmFormer and ShaSpec (Zhang et al., 2022; Wang et al., 2023) improve prediction when inputs are absent. Recent reconstruction, distillation, and configuration-specific experts extend this line in brain-tumor MRI (Zeng et al., 2024; Kang et al., 2026; Zhu et al., 2025; Novosad et al., 2024). These methods optimize predictive robustness under a chosen exposure distribution. Our controlled family comparison instead asks whether the measured separation between predictors changes when that distribution changes. mmFormer's modality-token fusion and ShaSpec's shared–specific factorization provide distinct, directly adaptable fusion designs: each can receive an identical external masking schedule while retaining its published representation and loss.

Learning under missingness shift. Missing-data theory establishes that observation mechanisms can alter the target estimand (Rubin, 1976; Little & Rubin, 2002). DAMS studies source–target adaptation when missingness mechanisms change (Zhou et al., 2023); StableMiss and StableMiss+ learn predictors for agnostic mask-distribution shift (Zhu et al., 2023; Zhu & Jiang, 2024); and recent analysis characterizes when Bayes predictors remain stable under ignorable missingness shifts (Rockenschaub et al., 2024). Shades-of-Null evaluates imputation quality, predictive performance, stability, and fairness across mixed and shifted missingness mechanisms (Khan et al., 2025). These works target predictor adaptation, imputation, or modality-value estimation. OCS instead conditions on fitted predictors and asks how a method-to-method margin responds when the benchmark changes its full joint distribution over observed MRI sequences. Training-exposure matching is then a controlled second test: it strengthens each comparator under the workflow prior without changing the estimand or family architecture.

Comparative sensitivity. Pipeline variance, task choice, and test-set recollection can alter reported performance and rankings (Bouthillier et al., 2021; Dehghani et al., 2021; Recht et al., 2019). Here the varied object is the MRI observability distribution, with splits, preprocessing, fitted ensembles, paired patients, and mask realizations controlled within each OCS comparison. The resulting statistic $[S_r(a) - S_r(b)] - [S_{\mathrm{Rand}(0.5)}(a) - S_{\mathrm{Rand}(0.5)}(b)]$ removes score movement common to both methods and measures the remaining change in competitive separation. Standard robustness curves vary a scalar missing rate for one model; OCS changes the unit of analysis to a paired method conclusion under a frozen joint intervention. The primary Protocol contrast evaluates the complete empirical acquisition mixture—its modality marginals, visible-cardinality distribution, support, and cross-modality dependence—because that mixture is the benchmark object encountered by the comparison. Center- and time-specific priors test sensitivity to how that object is estimated, matched exposure

tests whether the response is a training recipe artifact, and development-isolated centers separately validate external segmentation performance under naturally occurring acquisition mixtures.

3 OBSERVABILITY-INDEXED MODEL COMPARISON

3.1 A FROZEN JOINT INTERVENTION

The benchmark variable is a distribution over the four-sequence acquisition vector rather than a scalar missing rate. Write $\mathbf{z} \in \{0,1\}^4$ in the fixed order (T1, T1Gd, T2, FLAIR), where one denotes an acquired sequence. Rand(0.5) assigns $p_{\mathrm{Rand}}(\mathbf{z}) = 2^{-4}$ to each of the sixteen states, including the all-missing state. Protocol uses the unsmoothed empirical table

$$\hat{p}_{\mathrm{Protocol}}(\mathbf{z}) = \frac{n_{\mathrm{A/B}}(\mathbf{z})}{9{,}842}, \tag{1}$$

estimated once from Centers A/B metadata and held fixed across model families. Each metadata-log entry contributes one binary vector; no image, segmentation label, model output, or external-center record enters the estimate. This construction makes modality availability $\sum_{\mathbf{z}:z_m=1} \hat{p}(\mathbf{z})$, visible cardinality $\sum_m z_m$, empirical support, and cross-modality dependence explicit properties of the intervention. LOWQ is introduced after the binary draw by the separate degradation process in Section 4. Protocol therefore tests the complete workflow-derived acquisition mixture rather than attributing a response to any one of its distributional summaries.

3.2 MARGIN RESPONSE AND EXTERNAL VALIDATION

Training and evaluation observability have different indices. Let q_a denote the masking exposure used to fit method a, and let $S_r(a \mid q_a)$ be its score when the resulting fixed model is evaluated under regime r. For methods a and b, the evaluation-regime margin is

$$M_r(a, b \mid q_a, q_b) = S_r(a \mid q_a) - S_r(b \mid q_b). \tag{2}$$

For a prespecified alternative r, *observability conclusion sensitivity* is

$$\mathrm{OCS}_r(a, b \mid q_a, q_b) = M_r(a, b \mid q_a, q_b) - M_{\mathrm{Rand}(0.5)}(a, b \mid q_a, q_b). \tag{3}$$

Every OCS comparison holds both fitted ensembles fixed and changes only their evaluation masks. Within each regime, methods are paired by patient and mask realization; the subtraction across regimes then produces the difference in paired margins. The primary statistic uses $r = \mathrm{Protocol}$; protocol-aware comparator rows set $q_a = q_b = \mathrm{Protocol}$ and are reported separately from common-pipeline rows. OCS is the interaction between method identity and the named full evaluation distribution, expressed on the same effect-size scale as the leaderboard.

At external center c, the clean-transfer margin $\mathrm{CTM}_c(a, b) = S_c(a \mid q_a) - S_c(b \mid q_b)$ is reported separately for Centers C and D. Here "clean" denotes exclusion from training, prior estimation, hyperparameter choice, and model selection. The conclusion-robustness profile comprises the reference margin, Protocol margin, paired OCS interval, and center-specific CTMs. OCS quantifies the controlled benchmark intervention, while CTM provides complementary external performance evidence under each center's natural acquisition mixture.

3.3 A STATE-AWARE TEST INSTRUMENT

The four MRI sequences {T1, T1Gd, T2, FLAIR} form the set $\mathcal{M}$. For a modality $m \in \mathcal{M}$, an image volume x_m is available only if the modality is present; its acquisition state $s_m \in \{\mathrm{ACQ}, \mathrm{MISS}, \mathrm{LOWQ}\}$ indicates an acquired sequence, an absent sequence, or a present sequence judged to have low quality, respectively. We aim to predict WT, TC, and ET labels voxel-wise. Our central modeling assumption treats the modality-state vector $\mathbf{s}$ as a draw from the structured deployment distribution $p_{\mathrm{deploy}}(\mathbf{s})$, rather than sampling modalities independently; this distribution arises from acquisition protocols, workflow constraints, and quality failures. Standard missing-modality models consequently face three related failure modes: identity-only fusion cannot distinguish absence from degraded inputs, i.i.d. modality dropout mismatches deployment missingness, and prediction instability across admissible modality subsets degrades confidence estimates.

ProtocolProbe uses acquisition status directly as input to address the first failure mode. An identity embedding $\mathbf{e}_m^{\mathrm{id}} \in \mathbb{R}^d$ is assigned to each modality, while a state embedding $\mathbf{e}_{s_m}^{\mathrm{st}} \in \mathbb{R}^d$ is assigned to each state. Adding these embeddings yields the modality token, $\mathbf{t}_m = \mathbf{e}_m^{\mathrm{id}} + \mathbf{e}_{s_m}^{\mathrm{st}}$. When a modality is acquired or has low quality, the shared encoder f_θ produces P patch tokens from its image, $\mathbf{Z}_m = f_\theta(x_m) \in \mathbb{R}^{P\times d}$; each patch then receives the state-aware token through addition: $\widetilde{\mathbf{Z}}_m = \mathbf{Z}_m + \mathbf{1}\mathbf{t}_m^\top$. Missing modalities receive no imputed intensities. Only their state-aware token enters the fusion stage. When all four modalities are missing, the fusion input therefore contains four MISS tokens and no image-derived patch tokens; zero-valued skip tensors preserve the decoder interface. Unlike identity-only embeddings, the design distinguishes "not acquired" from "acquired but unreliable," as required by the deployment problem (Table 5).

The input to fusion Transformer g_ϕ concatenates every present-modality patch set with the missing-modality tokens; a U-Net decoder with skip connections decodes the resulting fused representation. Adding soft Dice to cross-entropy gives the segmentation loss. Protocol-conditioned modality masking (PCMM) handles the train–test mismatch in $\mathbf{s}$ by drawing modality-state patterns using the same frozen empirical prior $\hat{p}(\mathbf{s})$, whose estimate uses modality-presence metadata from Centers A/B (Sec. 4); training thus encounters site-realistic correlations in acquired/missing patterns instead of independent dropout. This empirical prior does not generate the LOWQ state: a separate corruption process takes a sampled acquired/missing pattern and labels a subset of its acquired modalities LOWQ (Sec. 4). To handle the third failure mode, subset-consistency regularization (SCR) draws two admissible state patterns for a single case and uses symmetric KL divergence to penalize disagreement in their softmax outputs. Adding $\lambda\mathcal{L}_{\mathrm{SCR}}$ to the segmentation loss gives the complete training objective. Together, state tokens, PCMM, and SCR couple the representation, training distribution, and stability objective to the same explicit observability variable. The ID-only control removes this bundle while preserving the backbone; its OCS isolates the bundle-level response, while component ablations quantify contributions to Protocol score and calibration. Protocol-aware comparator adaptation applies the masking exposure to mmFormer and ShaSpec without changing their published fusion designs.

All segmentation and uncertainty metrics first average voxel probabilities within a $K{=}5$ independently initialized ensemble. Three complete train-and-evaluate repeats (seeds 0/1/2) then provide repeat-specific scores and paired case effects. Within a repeat and regime, every method receives the same pre-generated mask for a patient; the fitted checkpoint and mask stream remain fixed throughout resampling. Headline point estimates average the three repeat-specific estimates. For each of 1,000 patient bootstrap resamples, we compute the paired margin and OCS within every repeat and average the three replicate statistics; the 2.5th and 97.5th percentiles form the reported two-sided 95% interval. The directional primary p_{OCS} is the finite-sample-corrected bootstrap probability of a non-positive OCS. Mean$\pm$std in score tables separately exposes variation among the three full training repeats. Case uncertainty is the mean predictive entropy over predicted tumor voxels, with the top-50 tumor-probability voxels used when the post-processed mask contains fewer than 50 voxels. Appendix A.1 evaluates this ranking signal across multiple estimators, failure thresholds, and review budgets.

Appendix Figure 2 summarizes the information flow from acquisition state through encoding, fusion, and decoding.

4 Evaluation Design

Each experiment targets a distinct explanation for a changed model margin. The paired Rand–Protocol comparison measures the benchmark response with fitted ensembles fixed; exposure-matched comparators test whether unequal training masks explain it; separately frozen priors test estimation choices; fixed-backbone and component controls test the state-aware instrument; and Centers C/D evaluate external performance without development reuse. Preprocessing, patient splits, model selection, post-processing, and within-regime missingness realizations are common to the relevant method pair. Full-input performance checks capacity, while the paired OCS and development-isolated center margins carry the primary evidence.

4.1 DATASETS AND OBSERVABILITY INTERVENTIONS

All segmentation models train only on BraTS 2021 images and labels (Menze et al., 2015; Baid et al., 2021). The dataset contains 1,251 labeled T1, T1Gd, T2, and FLAIR studies with WT/TC/ET masks. We use a fixed 80/20 patient split for every method and a common validation subset within the development partition; the held-out partition contains 250 patients. The first external cohort comprises 312 retrospective glioma cases from Centers A/B/C (134/98/80), acquired on Siemens Prisma 3T, GE MR750 3T, and Philips Ingenia 1.5T systems under a unified labeling protocol. Centers A/B provide modality-presence metadata for prior estimation, whereas Center C contributes no image, metadata, hyperparameter, or model-selection signal. Appendix Table 6 summarizes image and metadata roles without pooling their denominators.

Full exposes all four sequences. Rand(0.5) drops each modality independently with probability 0.5 over all 2^4 binary states. The all-missing draw is retained with probability $1/16$ and processed through each architecture's missing-input path; it is never rejected or resampled. For ProtocolProbe, this path contains four MISS tokens, no image tokens, and zero-valued encoder skips (Section 3). Protocol samples acquired/missing vectors from $\hat{p}(\mathbf{s})$, estimated once from 9,842 de-identified Centers A/B modality-presence logs collected during 2019–2024. The sampling unit is a metadata-log encounter rather than a unique patient; distinct repeat encounters remain separate observations of the acquisition workflow. Eligible entries have resolvable acquired/missing status for each target sequence class. Pre-contrast T1, post-contrast T1, T2-weighted, and FLAIR series map to T1, T1Gd, T2, and FLAIR, respectively; multiple series in one class collapse to one acquired indicator. Each retained entry contributes one four-sequence binary vector; observed pattern counts are normalized into an unsmoothed discrete distribution and frozen before fitting, so its support is exactly the set of observed patterns. The primary table uses this full A/B prior. Sensitivity analysis re-estimates the same object from A-only (5,318 logs), B-only (4,524), 2019–2021 (4,327), and 2022–2024 (5,515) windows. The Laplace variant adds one pseudocount to each of the 2^4 states, including all-missing. For perturbations, $d = \arg\max_j \hat{p}_j$ is the all-four-acquired state in the A/B table; we set $\hat{p}'_d = (1+\delta)\hat{p}_d$ for $\delta \in \{-0.10, 0.10\}$ and multiply every $j \neq d$ by $(1-\hat{p}'_d)/(1-\hat{p}_d)$. Each variant is independently frozen and shared across methods. For SCR, both subset draws use the same admissible support as the corresponding training prior. At evaluation, each fitted ensemble receives both Rand and Protocol masks. Each repeat uses one pre-generated draw per patient and regime; within a regime and method pair, patient-level draws are identical. The OCS subtraction couples the two regime-specific paired margins, while score-table standard deviations expose variation across the three independent training and mask seeds.

The LOWQ state is applied after acquired/missing sampling by corrupting 30% of acquired sequences with motion blur ($\sigma \in [1, 3]$), $2\times$ downsampling, or bias fields. Mixed combines this degradation with Protocol missingness. Appendix A.1 also evaluates 84 external studies with blinded sequence-quality labels, covering motion, resolution loss, and bias-field degradation.

4.2 EXPOSURE-MATCHED COMPARATOR ADAPTATION

The common-pipeline comparison includes Zero-fill and Dropout Swin UNETR, HeMIS, U-HVED, mmFormer, and ShaSpec (Havaei et al., 2016; Dorent et al., 2019; Hatamizadeh et al., 2022; Zhang et al., 2022; Wang et al., 2023). The ID-only control retains ProtocolProbe's backbone and optimizer while removing acquisition-state tokens, PCMM, and SCR. Table 4 records implementation provenance.

For protocol-aware parity, mmFormer and ShaSpec retain their published encoders, fusion modules, and losses while receiving the same frozen prior at the model input. Their training loaders draw identical missingness realizations and the same number of masked examples per epoch as ProtocolProbe. All three families use the same optimizer-selection budget, validation subset, seeds 0/1/2, fresh $K{=}5$ ensembles, and 300-epoch schedule; the selected checkpoint maximizes WT/TC/ET-averaged validation Dice. No acquisition-state tokens or SCR terms are inserted into comparator architectures; parity therefore equalizes observability exposure without converting the families into ProtocolProbe. Their all-missing paths remain the family-specific missing-input representations used in the corresponding common-pipeline rows.

4.3 DEVELOPMENT-ISOLATED CENTER TRANSFER

Center D is an independent retrospective referral cohort of 96 pre-treatment adult glioma studies acquired on 1.5T and 3T systems with the same four-sequence target and unified WT/TC/ET labeling protocol. Eligibility requires a pathologically confirmed diffuse glioma, at least one diagnostic sequence, and a complete reference segmentation; post-treatment and non-diagnostic studies are excluded. Its mutually exclusive acquisition patterns contain 54 all-four, 21 no-contrast, 7 no-T1, 5 no-T2, 6 fast-protocol, and 3 FLAIR-only cases. Center D contributes no training image, prior-estimation metadata, hyperparameter decision, or model-selection signal.

4.4 IMPLEMENTATION AND STATISTICS

Volumes are resampled to 1 mm^3, skull-stripped with HD-BET (Isensee et al., 2019), and normalized per channel. ProtocolProbe uses a Swin UNETR encoder with depths 2-2-2-2 and window size 7, a four-layer fusion Transformer with eight heads and token dimension 384, and a U-Net decoder. Training uses PyTorch 2.1 (Paszke et al., 2019) and MONAI 1.3 (Cardoso et al., 2022) on four NVIDIA A100 80 GB GPUs (CUDA 12.1, cuDNN 8.9, driver 535.104.05). AdamW (Loshchilov & Hutter, 2019) uses learning rate 2×10^{-4}, weight decay 10^{-5}, a cosine schedule (Loshchilov & Hutter, 2017), 300 epochs, 128^3 patches, batch size two per GPU, and $\lambda = 0.5$ for SCR. Each run takes approximately 18 hours. Post-processing retains the largest connected component per subregion and removes components below 50 voxels; no test-time augmentation is used. The $K{=}5$ ensemble has 89.2M backbone and 0.3M token parameters and requires 3.2 s per volume on one A100.

We report WT/TC/ET-averaged Dice and HD95, plus ECE with 15 bins over tumor voxels and max-class probability. Means and standard deviations are computed over three fresh train-and-evaluate repeats (seeds 0/1/2). Within each repeat, five-model probabilities are averaged before case-level metrics and paired method differences are computed. For each of 1,000 patient resamples, the three repeat-specific paired statistics are then averaged; percentile bounds give two-sided 95% intervals. The reported primary p-value is the one-sided, finite-sample-corrected bootstrap tail probability for the prespecified positive OCS direction. Methods share the same patient and within-regime missingness realization in every pair. The full-A/B, ProtocolProbe-versus-PA-mmFormer OCS is the prespecified primary contrast. The seven alternate priors form a structured sensitivity envelope; their intervals are pointwise summaries, while inference is anchored to the prespecified full-A/B contrast. Failure detection uses case-level mean Dice thresholds 0.65/0.70/0.75, and Appendix A.1 reports AUROC, AUPRC, ECE, raw failure counts, and capture across 5/10/15/20/25% review budgets.

5 RESULTS

The results progress from the observed margin response to increasingly strict controls: exposure matching, prior construction, external performance, mechanism, and failure ranking.

5.1 JOINT OBSERVABILITY CHANGES COMPARATIVE SEPARATION

Against mmFormer, ProtocolProbe scores 79.3 versus 76.9 under Rand(0.5), a 2.4-Dice margin, and 78.9 versus 74.6 under Protocol, a 4.3-Dice margin. The resulting OCS is $+1.9$ Dice (paired 95% CI [0.9,3.0], one-sided bootstrap $p = 0.001$). Protocol ECE simultaneously improves from 0.046 to 0.027. ProtocolProbe is the highest-scoring row under both regimes, but the full leaderboard is not rank invariant: ShaSpec changes order relative to ID-only and U-HVED. OCS therefore targets a named method pair and its effect size rather than reducing benchmark stability to a top-row indicator.

5.2 CROSS-FAMILY EXPOSURE MATCHING

Table 2 adapts mmFormer and ShaSpec to the same frozen prior and exposure budget. Protocol-aware mmFormer improves over its common-pipeline counterpart, reducing ProtocolProbe's Rand, Protocol, and Ctr-C margins to 1.1, 2.1, and 2.4. Its parity-controlled OCS is 1.0 Dice, with paired 95% CI [0.3, 1.7]. Protocol-aware ShaSpec produces the same positive direction with OCS = 1.4.

Table 1: Comparison on the BraTS 2021 held-out split and development-isolated Center C. The ProtocolProbe–mmFormer margin is 2.4 under Rand(0.5) and 4.3 under Protocol. Mean Dice averages WT/TC/ET. Across three train-and-evaluate repeats (seeds 0/1/2), each using $K=5$ independently initialized ensembles, we report mean±std. Calibration in the Protocol regime is reported as Protocol ECE. The same-backbone pair carries †; ID-only omits acquisition-state tokens, PCMM, and SCR.

Method	Full	Rand (0.5)	Protocol	Ctr-C	Protocol ECE ↓
Zero-fill Swin UNETR	87.1±0.3	70.2±1.3	67.3±1.9	61.9±2.4	0.073±0.008
HeMIS (Havaei et al., 2016)	85.9±0.5	73.4±0.9	70.8±1.5	65.8±2.0	0.065±0.007
U-HVED (Dorent et al., 2019)	86.4±0.4	74.8±0.8	72.4±1.2	67.6±1.7	0.057±0.006
ShaSpec (Wang et al., 2023)	86.8±0.5	75.6±0.9	71.9±1.3	66.9±1.9	0.052±0.006
Dropout Swin UNETR	87.1±0.3	76.1±0.7	73.8±1.1	69.0±1.6	0.049±0.005
mmFormer (Zhang et al., 2022)	87.3±0.4	76.9±0.8	74.6±1.0	71.3±1.6	0.046±0.005
ID-only tokens†	87.2±0.3	75.4±0.8	73.2±1.1	69.5±1.7	0.047±0.005
ProtocolProbe (matched probe)†	87.4±0.3	79.3±0.6	78.9±0.7	76.1±1.0	0.027±0.003

Table 2: Exposure-matched comparator test. Panel A reports mean Dice and Protocol ECE across three repeats; PA denotes protocol-aware training. Panel B reports ProtocolProbe's margins relative to comparator b and shows positive OCS after adaptation. Panel C consolidates paired OCS intervals for the prespecified PA-mmFormer comparison and the original-mmFormer diagnostic using 250 held-out patients and 1,000 patient bootstraps.

(A) Method	Full	Rand	Protocol	Ctr-C	ECE
mmFormer	87.3 ± 0.4	76.9 ± 0.8	74.6 ± 1.0	71.3 ± 1.6	0.046 ± 0.005
ShaSpec	86.8 ± 0.5	75.6 ± 0.9	71.9 ± 1.3	66.9 ± 1.9	0.052 ± 0.006
PA-mmFormer	87.4 ± 0.4	78.2 ± 0.7	76.8 ± 0.8	73.7 ± 1.3	0.035 ± 0.004
PA-ShaSpec	86.9 ± 0.5	76.9 ± 0.8	75.1 ± 1.0	70.8 ± 1.6	0.041 ± 0.005
ID-only	87.2 ± 0.3	75.4 ± 0.8	73.2 ± 1.1	69.5 ± 1.7	0.047 ± 0.005
ProtocolProbe	87.4 ± 0.3	79.3 ± 0.6	78.9 ± 0.7	76.1 ± 1.0	0.027 ± 0.003

(B) Comparator b	M_{Rand}	M_{Prot}	M_{CtrC}	OCS
mmFormer	2.4	4.3	4.8	1.9
ShaSpec	3.7	7.0	9.2	3.3
PA-mmFormer	1.1	2.1	2.4	1.0
PA-ShaSpec	2.4	3.8	5.3	1.4
ID-only (probe contrast)	3.9	5.7	6.6	1.8

(C) Comparator b	M_{Rand}	M_{Prot}	OCS [95% CI]	Patients
mmFormer	2.4	4.3	1.9 [0.9,3.0]	250
PA-mmFormer	1.1	2.1	1.0 [0.3,1.7]	250

Both paired OCS intervals exclude zero; the original-mmFormer diagnostic has one-sided bootstrap $p = 0.001$. Protocol-aware retraining strengthens the comparators, while the ProtocolProbe–PA-mmFormer margin still increases under Protocol evaluation.

5.3 Sensitivity to prior construction

Figure 1 repeats the parity-controlled comparison for center-specific, temporal, smoothed, and perturbed priors. Every construction yields positive Protocol, Ctr-C, and OCS margins, and each pointwise paired OCS interval excludes zero. Protocol margins vary from 1.7 to 2.6, Ctr-C margins from 2.0 to 3.0, and OCS from 0.7 to 1.3. Prior composition changes the magnitude but not the observed direction.

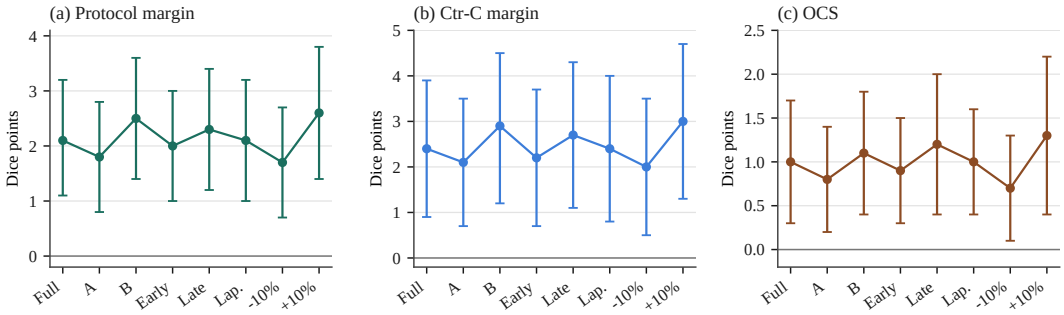


Figure 1: Paired-margin profile for ProtocolProbe versus protocol-aware mmFormer across eight separately frozen observability priors. Points are method margins; bars are paired patient-level 95% pointwise bootstrap intervals. Every OCS interval remains positive while the magnitude varies with prior composition. Protocol uses 250 held-out patients and Ctr-C uses 80. The full numeric matrix and prior denominators appear in Table 7.

5.4 Development-isolated external performance

At Center C, ProtocolProbe reaches 76.1 versus 71.3 for mmFormer, a 4.8-Dice margin with bootstrap 95% CI [3.0, 6.7] and $p < 0.001$. Against protocol-aware mmFormer, the Ctr-C margin is 2.4 (95% CI [0.9, 3.9]). Center D is harder in absolute Dice yet retains a 2.4-point parity-controlled margin (95% CI [0.8, 4.1], $p = 0.006$). These center results validate external performance under each center's natural acquisition mixture; OCS remains the paired Protocol-versus-Rand benchmark response on the held-out BraTS cohort.

Table 3: External margins at two development-isolated centers. Values are mean Dice±std across seeds 0/1/2 with fresh K=5 ensembles. The parity-controlled margin remains 2.4 Dice at both centers with distinct paired intervals. Neither center contributes images or metadata to training, prior estimation, or model selection.

Method	Center C ($n = 80$)	Center D ($n = 96$)
mmFormer	71.3±1.6	69.8 ± 1.8
PA-mmFormer	73.7 ± 1.3	72.0 ± 1.4
PA-ShaSpec	70.8 ± 1.6	69.9 ± 1.7
ID-only	69.5±1.7	68.6 ± 1.9
ProtocolProbe	76.1±1.0	74.4 ± 1.2
ProtocolProbe − PA-mmFormer	2.4	2.4
Paired 95% CI	[0.9,3.9]	[0.8,4.1]

The pooled Ext. A/B/C score is 75.6 for ProtocolProbe and 71.0 for mmFormer. Its aggregate can be slightly lower than Ctr-C because A/B add protocol and site heterogeneity; Ctr-C is development-isolated and measures a distinct transfer setting.

5.5 Fixed-backbone and component evidence

The ID-only control supplies a bundle-level gap-of-gaps test using Table 1. Relative to mmFormer, its margin changes from -1.5 under Rand(0.5) to -1.4 under Protocol, an OCS of $+0.1$; ProtocolProbe's corresponding OCS is $+1.9$. Their $+1.8$ difference is exactly the ProtocolProbe-versus-ID-only OCS in Panel B of Table 2. The shared backbone alone does not reproduce the observability-conditioned movement. Because ID-only removes state tokens, PCMM, and SCR jointly, this contrast attributes the difference to the complete bundle rather than to an individual component.

Appendix Table 5 reports the complete one-at-a-time Protocol and Mixed ablations. Removing acquisition-state tokens gives the largest Dice and ECE deterioration. Removing PCMM lowers Dice by 2.7 points; removing SCR raises ECE from 0.027 to 0.041 and lowers AUROC from 0.847 to 0.823.

In Mixed, removing LOWQ tokens lowers Dice by 2.4 points, consistent with the natural-quality validation in Table 8.

5.6 CALIBRATION AND FAILURE RANKING

Under Protocol, the exposure-matched comparison improves ECE from 0.035 ± 0.004 for PA-mmFormer to 0.027 ± 0.003 for ProtocolProbe and failure-detection AUROC from 0.801 ± 0.023 to 0.847 ± 0.018. At a 15% review budget, the same ensemble-entropy score captures 68.7% of ProtocolProbe's Dice< 0.70 failures versus 60.0% for PA-mmFormer. The corresponding median raw failure counts are 32 and 35; Appendix Table 9 reports the paired operating ledger. Appendix Tables 9 and 10 show that the ordering persists across five estimators, three failure thresholds, and five review budgets.

6 DISCUSSION

The central result changes what a stable leaderboard can certify. ProtocolProbe is the highest-scoring row under both Rand and Protocol, yet its margin over mmFormer changes by 1.9 Dice and the internal order of other rows also moves. After equalizing protocol exposure, the remaining 1.0-Dice OCS shows that the benchmark distribution still changes the effect size assigned to the same pair. Such movement affects claims of practical equivalence, expected gains, and sample-size planning even when a binary winner label is unchanged.

The evaluation design yields an actionable reporting unit for incomplete-modality benchmarks: name and freeze the joint acquisition distribution, separate training exposure from evaluation exposure, share masks within each paired comparison, report the reference and target margins with OCS, and reserve untouched centers for external performance validation. The eight-prior envelope shows why the distribution itself must be reported rather than summarized by one dropout rate, while the fixed-backbone control and protocol-aware comparators localize the response beyond backbone capacity or unequal mask exposure. The calibration analysis adds an operational consequence: at a 15% review budget, ensemble entropy captures 68.7% of ProtocolProbe failures versus 60.0% for exposure-matched mmFormer.

The evidence applies to retrospective WT/TC/ET-averaged segmentation, the reported mmFormer and ShaSpec controls, and the evaluated MRI acquisition mixtures. The primary intervention changes modality marginals, visible cardinality, support, and dependence together, so OCS identifies the response to that complete mixture; the seven alternate-prior intervals are pointwise sensitivity summaries, and the paired intervals condition on the fitted ensembles and frozen mask streams while repeat variability is reported separately. These boundaries keep the interpretation aligned with the experiment, while two development-isolated centers, natural low-quality studies, and center- and time-specific priors establish the result across the observability variation represented in this study.

7 CONCLUSION

Joint MRI acquisition patterns are part of the evidence produced by an incomplete-modality benchmark. OCS makes this dependence measurable by comparing the same fixed-model margin under two frozen observability distributions, and ProtocolProbe provides a state-aware instrument for the test. The paired response remains 1.0 Dice after protocol exposure is matched (95% CI [0.3, 1.7]), keeps a positive 0.7–1.3 range across eight prior constructions, and is accompanied by 2.4-Dice external performance margins at both development-isolated centers. Fixed-backbone, component, natural-quality, calibration, and failure-ranking analyses connect that benchmark response to explicit acquisition state and downstream error review. Incomplete-modality comparisons should therefore treat the full joint mask distribution as a reported experimental variable and pair every headline score with the margin response it induces.

AI USE STATEMENT

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed all edits and remain responsible for the final content.

REFERENCES

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C. Kitamura, Sarthak Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. URL https://arxiv.org/abs/2107.02314.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021. URL https://proceedings.mlsys.org/paper/2021/hash/0184b0cd3cfb185989f858a1d9f5c1eb-Abstract.html.

M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. URL https://arxiv.org/abs/2211.02701.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021. URL https://arxiv.org/abs/2107.07002.

Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 74–82. Springer, 2019. doi: 10.1007/978-3-030-32245-8_9.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, pp. 272–284. Springer, 2022. doi: 10.1007/978-3-031-08999-2_22.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9901 of *LNCS*, pp. 469–477. Springer, 2016. doi: 10.1007/978-3-319-46723-8_54.

Fabian Isensee, Marianne Schell, Irada Tursunova, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Andreas von Deimling, Klaus H. Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, 40(17):4952–4964, 2019. doi: 10.1002/hbm.24750.

Ming Kang, Fung Fung Ting, Shier Nee Saw, Raphaël C.-W. Phan, Zongyuan Ge, and Chee-Ming Ting. A multimodal feature distillation with mamba-transformer network for brain tumor segmentation with incomplete modalities. *IEEE Transactions on Artificial Intelligence*, 2026. doi: 10.1109/TAI.2026.3678550.

Falaah Arif Khan, Denys Herasymuk, Nazar Protsiv, and Julia Stoyanovich. Still more shades of null: An evaluation suite for responsible missing value imputation. *Proceedings of the VLDB Endowment*, 18(9):2899–2913, 2025. doi: 10.14778/3746405.3746416.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2 edition, 2002. doi: 10.1002/9781119013563.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016. doi: 10.1109/TPAMI.2015.2461544.

Philip Novosad, Richard A. D. Carano, and Anitha Priya Krishnan. A task-conditional mixture-of-experts model for missing modality segmentation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15009, pp. 34–43. Springer Nature Switzerland, October 2024. doi: 10.1007/978-3-031-72114-4_4.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Patrick Rockenschaub, Zhicong Xian, Alireza Zamanian, Marta Piperno, Octavia-Andreea Ciora, Elisabeth Pachl, and Narges Ahmidi. Robust prediction under missingness shifts. *arXiv preprint arXiv:2406.16484*, 2024. URL https://arxiv.org/abs/2406.16484.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887, 2023. doi: 10.1109/CVPR52729.2023.01524.

Zhilin Zeng, Zelin Peng, Xiaokang Yang, and Wei Shen. Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008, pp. 424–433. Springer Nature Switzerland, October 2024. doi: 10.1007/978-3-031-72111-3_40.

Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 107–117. Springer, 2022. doi: 10.1007/978-3-031-16443-9_11.

Helen Zhou, Sivaraman Balakrishnan, and Zachary Lipton. Domain adaptation under missingness shift. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9577–9606. PMLR, 2023. URL https://proceedings.mlr.press/v206/zhou23b.html.

Shenghao Zhu, Yifei Chen, Weihong Chen, Yuanhan Wang, Chang Liu, Shuo Jiang, Feiwei Qin, and Changmiao Wang. Bridging the gap in missing modalities: Leveraging knowledge distillation and style matching for brain tumor segmentation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15967, pp. 95–106. Springer Nature Switzerland, September 2025. doi: 10.1007/978-3-032-04984-1_10.

Yichen Zhu and Bo Jiang. StableMiss+: Prediction with incomplete data under agnostic mask distribution shift. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7190–7194. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446980.

Yichen Zhu, Jian Yuan, Bo Jiang, Tao Lin, Haiming Jin, Xinbing Wang, and Chenghu Zhou. Prediction with incomplete data under agnostic mask distribution shift. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4720–4728. International Joint Conferences on Artificial Intelligence Organization, 2023. doi: 10.24963/ijcai.2023/525.

ETHICS STATEMENT

The external cohort is retrospective and de-identified, and it was used under IRB approval and data-use agreements. The evaluated system is intended for clinical decision support with human review.

REPRODUCIBILITY STATEMENT

Section 3 specifies the paired-margin estimand, training and evaluation indices, model, and uncertainty procedure; Section 4 reports splits, comparator adaptation, hardware, optimization, masking, and evaluation conventions. Appendix A.1 provides cohort counts, bootstrap procedures, and supporting analyses.

A.1 SUPPORTING PROTOCOL AND RESULTS

A.1.1 COMPARATOR PROVENANCE AND PARITY CONTROLS

All rows use the same BraTS 2021 split, preprocessing, validation rule, missingness realization, and post-processing. Table 4 records the implementation path and the exact protocol-aware intervention applied to each family.

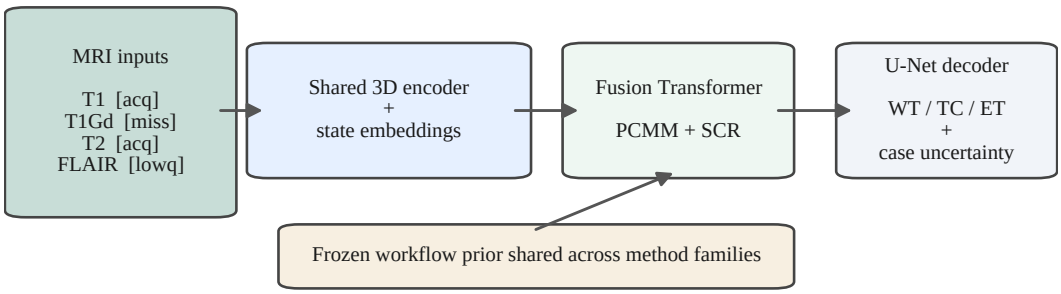


Figure 2: ProtocolProbe information flow. Each MRI sequence is paired with an explicit ACQ, MISS, or LOWQ state. Present sequences pass through a shared encoder, missing sequences contribute tokens only, and the fusion Transformer conditions prediction on identity and state. PCMM and SCR share the frozen observability object used by the cross-family intervention.

Table 4: Comparator provenance under the shared stress-test pipeline. PA rows retain the published family architecture while matching the frozen prior, exposure count, optimization budget, seeds, and evaluation.

Row	Starting point	Matched intervention
Zero-fill / Dropout Swin UNETR	Maintained MONAI Swin UNETR path	Common preprocessing and evaluation; zero filling or i.i.d. modality dropout provides the control intervention.
HeMIS / U-HVED (Havaei et al., 2016; Dorent et al., 2019)	Published architectures reimplemented in MONAI	Published losses with shared split, masking generator, validation, and post-processing.
mmFormer / ShaSpec (Zhang et al., 2022; Wang et al., 2023)	Public reference implementations	Common-pipeline rows preserve their training objectives; PA rows additionally receive the frozen prior and matched masking exposure.
ID-only	ProtocolProbe backbone and optimizer	Removes acquisition-state tokens, PCMM, and SCR while preserving the remaining training and evaluation path.
ProtocolProbe	Swin UNETR with state-aware fusion	Couples acquisition-state tokens, PCMM, and SCR to the frozen prior.

All-missing convention. Rand(0.5) retains the $(0, 0, 0, 0)$ state with probability $1/16$. It is neither rejected nor resampled. ProtocolProbe receives four MISS tokens, no image-derived patch tokens, and zero-valued skip tensors; ID-only follows the corresponding identity-token path. Every other row retains its family-specific missing-input representation in both the common-pipeline and PA variants, so adaptation changes mask exposure rather than the architecture's empty input semantics.

A.1.2 COMPONENT ABLATIONS

Table 5: Component ablations on BraTS 2021. Each row removes one component from the complete probe; all variants share the backbone and use three repeats with fresh $K=5$ ensembles. Protocol contains acquired/missing states, while Mixed adds degraded acquired sequences.

Configuration	Dice↑	HD95↓	ECE↓	AUROC↑
All components	78.9±0.7	9.6±1.4	0.027±0.003	0.847±0.018
− acq. tokens	74.1±1.0	13.8±2.1	0.048±0.005	0.791±0.022
− PCMM (rand. drop)	76.2±0.9	11.4±1.7	0.039±0.004	0.812±0.021
− SCR	77.4±0.8	10.7±1.5	0.041±0.004	0.823±0.019
− lowq tokens	77.8±0.7	10.2±1.4	0.034±0.004	0.831±0.020
Mixed (miss+lowq):				
All components	74.8±0.8	11.2±1.6	0.033±0.004	0.832±0.020
− lowq tokens	72.4±1.0	13.1±1.9	0.047±0.005	0.798±0.023

A.1.3 COHORT ROLES AND ISOLATION

Table 6 separates image cohorts from the workflow-log denominators used to estimate observability priors.

Table 6: Cohorts used by the reported evaluation. External MRI labels follow the unified WT/TC/ET reference protocol described in Section 4.

Source	MRI cases	Metadata logs	Role
BraTS 2021	1,251	–	Training/validation and the 250-patient primary held-out evaluation; four target sequences and WT/TC/ET masks.
Center A	134	5,318	A/B prior estimation and pooled external evaluation.
Center B	98	4,524	A/B prior estimation and pooled external evaluation.
Center C	80	0	Development-isolated external evaluation.
Center D	96	0	Independent development-isolated referral-cohort evaluation.

A.1.4 PAIRED STATISTICS AND LEAKAGE CONTROLS

Each reported repeat trains a fresh five-model ensemble at seeds 0/1/2. For a paired gap, voxel probabilities are first averaged within each ensemble, then method differences are computed on the same patient and missingness realization. Each of the 1,000 bootstrap replicates resamples patients with replacement while retaining these pairs, computes the margin and OCS within each training repeat, and averages the three repeat-specific statistics. The reported 95% interval uses the 2.5th and 97.5th percentiles of that distribution. The primary directional p_{OCS} uses $(1 + \#\{\mathrm{OCS}^* \leq 0\})/(1 + 1{,}000)$. Mean±standard deviation in score tables separately summarizes the three repeat-level point estimates. Center C and Center D intervals use the same patient-level procedure within each center. The intervals condition on the fitted ensembles and pre-generated mask streams; training-repeat variation remains visible in the reported repeat standard deviations. The full-A/B PA-mmFormer comparison is primary; alternate-prior intervals are pointwise sensitivity summaries. The A/B prior is estimated before model fitting, and Center C/D provide no training, prior-estimation, validation, or selection signal. Hyperparameters are shared across external centers and observability regimes.

A.1.5 FULL FROZEN-PRIOR SENSITIVITY MATRIX

Table 7 gives the numerical values underlying Figure 1. Protocol margins use 250 held-out BraTS patients; Ctr-C margins use 80 unseen-center patients. Every confidence interval uses 1,000 paired patient-level resamples.

Table 7: Complete paired-margin profile for ProtocolProbe relative to protocol-aware mmFormer. Log counts are the metadata denominator used to estimate each prior; smoothing and perturbations operate on all 9,842 logs. Intervals are pointwise paired 95% bootstrap intervals.

Frozen prior	Logs	Protocol margin [CI]	Ctr-C margin [CI]	OCS [CI]
Full A/B	9,842	2.1 [1.1,3.2]	2.4 [0.9,3.9]	1.0 [0.3,1.7]
A-only	5,318	1.8 [0.8,2.8]	2.1 [0.7,3.5]	0.8 [0.2,1.4]
B-only	4,524	2.5 [1.4,3.6]	2.9 [1.2,4.5]	1.1 [0.4,1.8]
2019–2021	4,327	2.0 [1.0,3.0]	2.2 [0.7,3.7]	0.9 [0.3,1.5]
2022–2024	5,515	2.3 [1.2,3.4]	2.7 [1.1,4.3]	1.2 [0.4,2.0]
Laplace	9,842	2.1 [1.0,3.2]	2.4 [0.8,4.0]	1.0 [0.4,1.6]
Dominant bin -10%	9,842	1.7 [0.7,2.7]	2.0 [0.5,3.5]	0.7 [0.1,1.3]
Dominant bin $+10\%$	9,842	2.6 [1.4,3.8]	3.0 [1.3,4.7]	1.3 [0.4,2.2]

A.1.6 External acquisition composition and performance

Figure 3 displays the exact, mutually exclusive acquisition-pattern counts for Center D and the exposure-matched external comparison at Centers C and D.

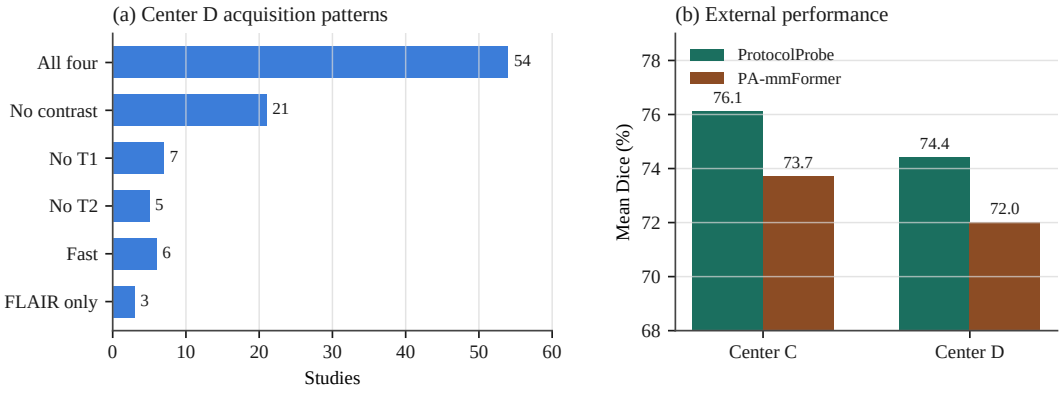


Figure 3: Development-isolated external acquisition and performance. Panel (a) reports all 96 Center D studies as an exhaustive partition of acquisition patterns. Panel (b) compares ProtocolProbe with PA-mmFormer at Centers C and D; neither center contributes training, prior-estimation, validation, or model-selection signal.

Center D contains 96 eligible studies: 54 all-four, 21 no-contrast, 7 no-T1, 5 no-T2, 6 fast-protocol, and 3 FLAIR-only acquisitions. All cases satisfy the inclusion criteria in Section 4; no Center D signal enters development.

A.1.7 Naturally degraded acquisition validation

Blinded quality review identifies 31 studies with motion, 28 with resolution loss, and 25 with dominant bias-field degradation. Table 8 evaluates each method on the resulting 84-study external subset.

Table 8: Natural low-quality acquisition results. Failures denote case-level mean Dice below 0.70; all metrics use the same blinded subset and paired predictions.

Method	Dice	HD95	ECE	Failures
ProtocolProbe	73.1 ± 1.3	12.4 ± 1.9	0.038 ± 0.005	14/84
PA-mmFormer	70.5 ± 1.6	15.1 ± 2.4	0.051 ± 0.006	20/84
ID-only	68.9 ± 1.8	16.7 ± 2.6	0.058 ± 0.007	24/84

ProtocolProbe exceeds PA-mmFormer by 2.6 Dice (paired 95% CI [0.8,4.5]). The wider metric dispersion relative to Protocol is consistent with the heterogeneous natural degradation types.

A.1.8 UNCERTAINTY ESTIMATORS AND OPERATING POINTS

Figure 4 summarizes the exposure-matched ensemble comparison against PA-mmFormer. Tables 9 and 10 extend the analysis to protocol-aware mmFormer, multiple estimators, thresholds, and budgets.

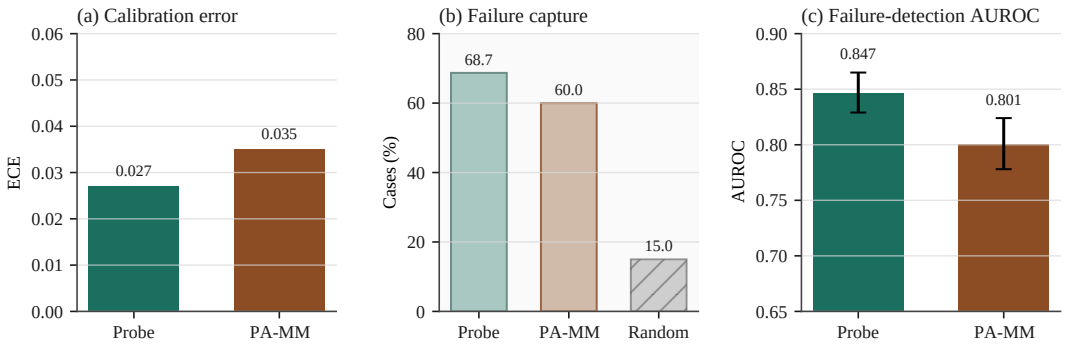


Figure 4: Protocol calibration and failure ranking. Panels report ECE, failure capture at a 15% review budget for Dice<0.70, and ensemble-entropy failure-detection AUROC. Probe denotes ProtocolProbe; PA-MM denotes protocol-aware mmFormer, and the random-review bar is the same-budget reference.

Table 9: Uncertainty estimator comparison at Dice<0.70 and a 15% review budget. Metrics are mean$\pm$std over three repeats; capture reports the corresponding identified-failure count.

Method	Score	ECE	AUROC	AUPRC	Captured
ProtocolProbe	Ensemble entropy	0.027 ± 0.003	0.847 ± 0.018	0.792 ± 0.020	22/32
ProtocolProbe	Single-model entropy	0.034 ± 0.004	0.812 ± 0.021	0.754 ± 0.024	19/33
ProtocolProbe	MC dropout	0.031 ± 0.004	0.826 ± 0.020	0.769 ± 0.022	20/33
ProtocolProbe	Max softmax	0.041 ± 0.005	0.783 ± 0.026	0.718 ± 0.029	17/33
ProtocolProbe	Temperature scaled	0.029 ± 0.003	0.806 ± 0.022	0.741 ± 0.025	18/33
PA-mmFormer	Ensemble entropy	0.035 ± 0.004	0.801 ± 0.023	0.742 ± 0.026	21/35
PA-mmFormer	Single-model entropy	0.043 ± 0.005	0.768 ± 0.028	0.701 ± 0.031	18/36
PA-mmFormer	MC dropout	0.040 ± 0.005	0.781 ± 0.025	0.719 ± 0.028	19/36
PA-mmFormer	Max softmax	0.052 ± 0.006	0.739 ± 0.031	0.661 ± 0.035	16/36
PA-mmFormer	Temperature scaled	0.038 ± 0.004	0.762 ± 0.027	0.693 ± 0.030	17/36

The ensemble score leads across thresholds, while MC dropout remains the strongest single-network alternative. ProtocolProbe preserves the higher capture rate at every evaluated budget, with non-uniform separation as the threshold changes.

Table 10: Ensemble-entropy failure capture across Dice thresholds and review budgets. Each cell is the repeat-averaged percentage of failures captured; $\widetilde{F}$ is the median raw failure count across the three repeats, so percentages are computed before aggregation.

Threshold	Method	$\widetilde{F}$	5%	10%	15%	20%	25%
0.65	ProtocolProbe	21	33.4	52.1	66.8	76.4	84.1
0.65	PA-mmFormer	24	25.0	41.7	54.2	66.7	75.0
0.70	ProtocolProbe	32	25.1	46.9	68.7	78.2	87.5
0.70	PA-mmFormer	35	20.0	37.2	60.0	71.4	80.0
0.75	ProtocolProbe	47	21.3	40.4	61.7	74.5	83.0
0.75	PA-mmFormer	53	17.0	34.0	52.8	66.0	75.5