



(RESEARCH ARTICLE)



STRESS TESTING SPATIAL ROUTING OF SYNTHESISED T1CE FOR MISSING-MODALITY BRAIN-TUMOUR SEGMENTATION USING TWO-DIMENSIONAL PROCEDURAL PHANTOMS

Zahraa S. Abd-Al Ameer *

Department of Physics, College of Science, Al-Nahrain University, Baghdad, Iraq.

* Corresponding Author

International Journal of Science and Research Archive, 2026, 20(03), 980–1001

Publication history: Received on 15 August 2026; revised on 24 September 2026; accepted on 26 September 2026

Article DOI: <https://doi.org/10.30574/ijrsra.2026.20.3.1805>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Apparent gains from adaptive fusion of synthesised contrast-enhanced T1-weighted (T1ce) MRI may reflect unequal label budgets, global calibration, or meta-model flexibility rather than genuinely spatial routing. We stress-tested this mechanism in two separately locked 16-seed studies using procedural phantoms and a 32-pair counterfactual control. Study 1 compared a restricted spatial router (QGate-MRI-R) with ungated synthesis-assisted segmentation, label-matched controls, and balanced logistic stacking. Study 2 introduced factorised acquisition shifts, alternative stackers, case-mean and histogram-preserving shuffled weights, and a target-only exposure test. In Study 1, QGate reduced in-distribution macro Brier loss versus ungated synthesis by 2.56×10^{-5} (Holm-adjusted $p = 0.0269$), but balanced logistic stacking performed better in distribution and synthesis with 20 additional labels achieved lower shifted Brier loss. None of the six Study 2 comparisons was significant after Holm adjustment; both locality tests had adjusted $p = 0.0551$. Across-seed variation in mean regional weights greatly exceeded within-map spatial variation, although the design did not separate fitting variability from regenerated case composition. Target-only exposure reduced PSNR by 12.22 dB while all 640 case-level invariance checks passed, and identical-observable pairs established construction-specific bounds of 0.25 for ET Brier loss on the ambiguity support and 2/3 for paired ET Dice. These findings support a mechanism stress test, not robust spatial-routing or clinical claims.

Keywords: Missing-Modality MRI; Brain-Tumour Segmentation; T1ce Synthesis; Spatial Routing; Distribution Shift; Procedural Phantoms; Identifiability

1 INTRODUCTION

Multimodal magnetic resonance imaging (MRI) provides complementary anatomical and pathological information for brain-tumour segmentation. T1-weighted imaging contributes structural detail, T2-weighted and fluid-attenuated inversion recovery (FLAIR) images emphasise water-rich abnormality and oedema, and contrast-enhanced T1-weighted imaging (T1ce) highlights enhancement associated with blood-brain-barrier disruption. This complementarity underlies the whole-tumour (WT), tumour-core (TC), and enhancing-tumour (ET) targets used in the Brain Tumor Segmentation benchmarks [1, 2]. A complete modality set is not always available, however. T1ce may be omitted because of restrictions on contrast administration, shortened protocols, motion-related rejection, incomplete

archival transfer, or differences between acquisition sites. Missingness can therefore remove task-relevant evidence while also introducing a distribution shift relative to the data used for model development.

Research on incomplete multimodal segmentation has produced several complementary strategies. Shared or modality-invariant representations allow a model to aggregate whichever inputs are present; hetero-modal variational models combine completion with segmentation; region-aware and transformer-based fusion methods preserve cross-modal structure; and masked representation learning exposes a model to missing inputs during training [3–11]. Explicit synthesis is another attractive strategy because it restores the expected input configuration, creates an image that can be inspected independently, and can support more than one downstream task [12–15]. Its weakness is error propagation: a generator can smooth a small enhancing focus, displace an edge, or create a plausible but incorrect local pattern. Global image metrics may change little even when such an error affects a small lesion or boundary substantially.

This distinction separates reconstruction fidelity from decision utility. Fidelity asks whether the generated image resembles the target under intensity or structural metrics such as mean absolute error, peak signal-to-noise ratio, or structural similarity. Utility asks whether using that image improves the downstream segmentation. The two are related but not interchangeable. A synthetic image can be visually convincing yet harmful to a classifier, while an imperfect image can retain a stable contrast cue that is useful to the task. Uncertainty propagation and adaptive fusion have consequently been proposed as safeguards [16–22]. These methods are important advances, but an uncertainty map is not automatically a map of downstream benefit: ensemble members may agree on the same systematic error, and a moderately uncertain region may still contain useful complementary information.

A spatial router is intended to address this problem by retaining an available-only branch and selectively mixing it with a synthesis-assisted branch. Nevertheless, an observed performance gain does not by itself demonstrate that useful local evidence was routed to the correct voxels. A routed model may have received more labels than its comparator, exploited correlations specific to one generator, or benefited from the flexibility of a second-level model. A visually detailed weight map may also function mainly as a case-level or region-global coefficient. Stacked generalisation is therefore an essential comparator because it can learn stable relations between branch outputs without requiring a spatial-locality interpretation [23]. Calibration work similarly shows that probability quality and thresholded overlap can rank the same methods differently [24–26]. Mechanism-level evaluation must consequently equalise label budgets, compare flexible and restricted fusion rules, and test the spatial arrangement of the weights rather than only their aggregate performance.

Procedural digital phantoms are useful for such tests because they provide exact labels and allow acquisition factors to change while anatomy is held fixed [27–30]. The same case can be exposed to one artefact at a time; a learned weight map can be replaced by its mean or shuffled while its histogram is preserved; and the unobserved target can be changed without altering the inputs available to a deployed missing-modality model. Phantoms also permit exact observable collisions in which two samples share identical T1, T2, and FLAIR inputs but differ in hidden enhancement. These operations support falsification tests that are difficult to construct retrospectively. Their scope is necessarily limited: procedural anatomy and authored artefacts cannot establish patient-level accuracy, scanner generalisation, safety, or clinical benefit.

The present work combines two separately locked procedural studies and one deterministic analytic control. The original 16-seed experiment compared QGate-MRI-R with ungated synthesis-assisted segmentation, equal and global fusion, synthesis-uncertainty weighting, balanced logistic stacking, additional-label controls, and descriptive test-label-selector and full-modality references under an in-distribution generator and a separately authored generator-B shift. Because its evidence was mixed, a separately locked 16-seed extension used paired factorised acquisition changes, stronger logistic comparators, and locality-removal controls. A T1ce-only exposure experiment tested whether reconstruction metrics could change while the complete deployable path remained invariant, and 32 exact identical-observable pairs established construction-specific limits when hidden ET evidence was absent from all available inputs. The two statistical studies retained separate Holm families and were not pooled as an effective sample of 32 seeds.

This study examined whether the original QGate result survived label-budget controls and generator shift; whether QGate exceeded ungated synthesis-assisted segmentation or a signed logistic regret blend under a separately locked compound acquisition shift; and whether native spatial weights improved Brier loss beyond case-mean or histogram-preserving shuffled weights. The exposure and counterfactual controls tested whether reconstruction fidelity is a sufficient surrogate for deployable utility and what limit follows when distinct hidden ET states produce the same observed modalities. The aim was an auditable mechanism test, not a universal fusion rule or clinically validated MRI algorithm.

2 MATERIALS AND METHODS

The investigation comprised three evidence layers. Study 1 was the completed original experiment, identified as MAX in the locked protocol, and was locked before its 16 final master seeds. Study 2 was a separately designed factorised extension, locked after software-only development runs and before 16 new final seeds. The third layer was an exact 32-pair counterfactual construction that verified deterministic bounds and was not part of either statistical family. One complete training seed, including regenerated training partitions, fitted models, calibration, and testing, was the independent unit for both inferential studies. The two 16-seed sets were analysed separately and never treated as a pooled $n = 32$ sample.

Table 1 defines the inferential role and admissible scope of each evidence layer.

Table 1: Evidence layers, inferential roles, and scope limits.

Layer	Evaluation material	Inference	Admissible scope
Study 1 (MAX protocol)	16 seeds; 640 ID + 640 OOD cases	12 Holm tests	Comparative behaviour under generators A/B
Factorised extension	16 new seeds; 40 cases $\times$ 13 conditions	6 Holm tests	Acquisition factors, locality controls, stronger stackers
Counterfactual	32 exact equal-prior pairs	No hypothesis test	Observable-collision bound for fixed input-only methods

The independent unit for both statistical layers was one complete training pipeline. Counterfactual pairs verified a theorem under a defined construction rather than estimating a population effect.

Figure 1 summarises the three evidence layers and their separation.

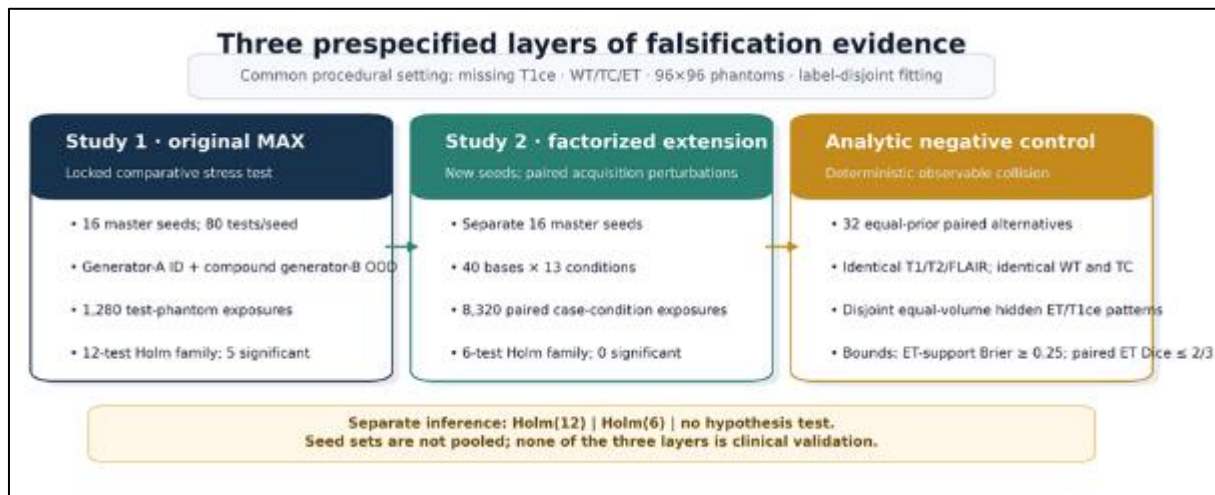


Figure 1: Three evidence layers and their non-pooled information flow. Study 1 and Study 2 used separate locks and separate sets of 16 training seeds. The 32-pair analytic control was outside both statistical families and was not treated as additional replication.

2.1 Procedural phantoms and data partitions

Each sample was a 96×96 two-dimensional brain-like phantom with a supplied brain mask, nested WT, TC, and ET labels, and aligned T1-, T2-, FLAIR-, and T1ce-like channels. Generator A varied ellipse-based anatomy, tumour geometry, tissue contrast, Rician-like noise, smooth multiplicative bias, blur, and weak ghosting. Generator B was separately authored and altered anatomy, topology, contrast, resolution, bias, motion, stripe, noise, and registration rules. It is described as an out-of-distribution (OOD) generator-B shift rather than an external cohort, site, or clinical validation set. The construction enabled exact labels and controlled pairing but did not model realistic three-dimensional anatomy, scanner physics, skull-stripping error, or patient heterogeneity.

Study 1 generated 40 locked generator-A in-distribution (ID) cases and 40 locked generator-B OOD cases per seed, with four balanced severities and ten cases per domain and severity. Study 2 instead generated 40 clean severity-zero

generator-A base cases per seed and copied each anatomy and label set across 13 prespecified conditions. These paired copies allowed individual artefacts and dropouts to be studied without changing the underlying case.

Per final seed, both studies regenerated 48 paired cases for synthesis training, 48 distinct cases for base segmentation, 12 cases for router or stacker fitting, and eight cases for calibration. The available-only and unconditional synthesis branches therefore used 48 segmentation-labelled cases, while QGate and supervised second-level models accessed 20 additional labelled cases. Additional-label controls were retrained on the same 48+20 segmentation-label budget. The separate 48 synthesis-training cases provided paired modality examples and were not counted as segmentation labels.

Table 2 details labelled-data access and the fairness role of every comparator family.

Table 2: Label budgets and comparator fairness.

Method family	Labelled cases per seed	Fairness role
Available only / synthesis	48 base	Frozen branch comparators
QGate-MRI-R	48 base + 12 meta + 8 calibration	68 labels across components
Balanced / natural stackers	48 + 12 + 8	Same second-level budget as QGate
Signed regret blend	48 + 12 + 8	Same budget; signed branch features
Available +20 / synthesis +20	68 base-equivalent	Label-matched retraining controls
Test-label selector / full anchor	Test labels / observed T1ce	Descriptive references only

QGate and every supervised stacker shared the same 12 meta and eight calibration cases. The +20 controls isolated the effect of additional segmentation supervision beyond the frozen 48-label branches.

2.2 Synthesis, segmentation, routing, and comparators

The missing-T1ce synthesiser was an extremely randomised tree regressor with 36 trees, maximum depth 18, minimum leaf size three, and maximum-feature fraction 0.78 [31]. Its mean prediction supplied synthetic T1ce, and normalised dispersion across trees supplied a tree-dispersion proxy rather than a validated uncertainty estimate. Available-only and synthesis-assisted segmenters used 52 trees, maximum depth 22, minimum leaf size two, maximum-feature fraction 0.80, balanced class weighting, and 70 sampled pixels per class per subject. QGate fitted separate WT, TC, and ET regret regressors with 64 trees, maximum depth 17, minimum leaf size eight, and maximum-feature fraction 0.82. Fourteen router features represented dispersion, entropy, confidence, unsigned disagreement and margins, binary disagreement, synthetic-image gradient and local standard deviation, and normalised dispersion. Direct signed branch probabilities, raw class vectors, raw synthetic intensity, local mean, and coordinates were excluded. Predicted regret was converted to a synthesis weight through a gain-restricted monotone sigmoid. Regional outputs were projected to preserve the ET-TC-WT nesting constraint, and final class labels were obtained by argmax.

Comparators included the two frozen branches, equal averaging, a calibration-set global convex blend, synthesis-dispersion weighting, balanced logistic stacking, additional-label available and synthesis controls, a test-label branch selector, and an observed-T1ce reference. The synthesis branch used without routing served as the ungated synthesis-assisted baseline. The extension added a natural-prevalence stacker and a signed logistic regret blend with the same 12 + 8 second-level labels as QGate. The test-label selector was an unattainable choice between only the available and synthesis branches and was not a universal ceiling. The observed-T1ce model was a full-information reference rather than a missing-modality method. QGate weights were continuous blend coefficients and were not interpreted as calibrated probabilities of benefit.

2.3 Factorised acquisition conditions and locality controls

Table 3 summarises the prespecified factorised acquisition conditions.

The 13 conditions were clean input; seven isolated artefacts comprising Rician noise, smooth bias, gamma contrast, anisotropic resolution, motion ghosting, periodic striping, and inter-modality misregistration; three single-modality dropouts; one compound available-input shift; and a T1ce-only exposure control. The locality pool contained only the seven one-factor artefacts. Each native weight map was compared with its case-specific regional mean and with a within-brain spatial permutation that preserved its histogram. Clean, compound, dropout, and exposure cases were excluded from the prespecified locality pool. The shuffle tested the implemented permutation and did not preserve spatial autocorrelation or smoothness. The exposure analysis failed if any observed input or asserted available, synthesis, QGate, or stacker prediction differed from clean.

Table 3: Prespecified factorised acquisition conditions.

Condition group	Prespecified construction
Clean	Unchanged severity-zero generator A
Single artefacts	Rician noise; smooth bias; gamma contrast; anisotropic resolution; motion ghosts; periodic stripes; intermodal misregistration
Dropouts	T1, T2, or FLAIR set to zero; descriptive only
Compound available shift	Moderate fixed combination of all seven single-artefact families
T1ce-only exposure	Observed T1/T2/FLAIR bitwise unchanged; only true unobserved T1ce translated

2.4 Outcomes, statistical analysis, and reproducibility

Segmentation outcomes were Dice, 95th-percentile Hausdorff distance (HD95), Brier loss, and 15-bin expected calibration error (ECE) for WT, TC, and ET, together with macro averages [33]. Reconstruction outcomes were brain-masked peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), mean absolute error (MAE), and root-mean-square error (RMSE) [32]. Router diagnostics included case-level average precision for the positive-regret indicator, regret-prediction error, mean regional synthesis weight, and within-map standard deviation. Average precision denotes the scikit-learn average_precision_score rather than trapezoidal area under a precision–recall curve.

Within each study, two-sided Wilcoxon signed-rank tests were applied to seed-level paired effects [34]. Study 1 controlled one 12-comparison family, and Study 2 controlled a separate six-comparison family by the Holm procedure [35]. Positive effects always favoured QGate: QGate minus comparator for Dice and comparator minus QGate for Brier. Confidence intervals were unadjusted 95% intervals from 10,000 seed-level bootstrap resamples [36], whereas inferential decisions used Holm-adjusted p values. Development seeds were excluded. Protocols, scripts, final seeds, output schemas, and multiplicity families were locked before final execution, and aggregation verified stored hashes and row counts before manuscript generation.

2.5 Identical-observable counterfactual control

The deterministic negative control generated 32 equal-prior pairs with bitwise-identical observed T1, T2, and FLAIR arrays, identical WT and TC masks, and two disjoint equal-volume ET supports encoded only in hidden T1ce. Any deterministic missing-T1ce method restricted to the observed inputs must return the same output for both members because it observes the same X. On the union of the disjoint supports, the equal-prior probability minimising Brier loss is 0.5, yielding an exact ambiguity-support lower bound of 0.25. For a shared deterministic hard prediction, set algebra gives a maximum paired-mean and worst-member ET Dice of 2/3. These construction-specific bounds do not apply to the true-T1ce reference and are not universal clinical ceilings.

3 RESULTS

3.1 Original locked experiment

All 16 final pipeline runs completed successfully, providing 640 ID and 640 generator-B OOD test phantoms. Five of the 12 prespecified comparisons were Holm-significant: two favoured QGate and three favoured a comparator. The full family is reported in Table 4, while Table 5 and Figure 2 summarise the principal macro outcomes and paired effects.

In-distribution macro Dice and macro Brier were 0.962602 and 0.00143570 for QGate, 0.962133 and 0.00146126 for ungated synthesis-assisted segmentation, 0.965062 and 0.00095991 for balanced logistic stacking, and 0.963021 and 0.00141031 for synthesis with 20 additional labels. Relative to ungated synthesis, QGate reduced ID macro Brier by 0.00002556 (unadjusted 95% CI 0.00000884 to 0.00004765; Holm-adjusted $p = 0.0269$), while its Dice increment did not survive correction. Balanced logistic stacking was better in distribution for both endpoints. Under generator B, QGate macro Dice and Brier were 0.823242 and 0.00729959. QGate exceeded the stacker only in shifted Dice by 0.044605 (Holm-adjusted $p = 0.0007$), whereas shifted Brier did not differ. Synthesis with 20 additional labels achieved lower shifted Brier than QGate by 0.0001685 (Holm-adjusted $p = 0.0242$). The original study therefore did not support general QGate superiority.

Table 4: Complete original 12-test primary family.

Domain	Comparator	Metric	Effect	95% CI	Holm p	Result
ID	Synthesis	Dice	+0.0004695	0.0001516 to 0.0008273	0.0916	Not significant
ID	Synthesis	Brier	+0.0000256	0.0000088 to 0.0000477	0.0269	QGate
ID	Balanced stacker	Dice	-0.0024597	-0.0035357 to -0.0012222	0.0131	Balanced logistic stacker
ID	Balanced stacker	Brier	-0.0004758	-0.0005127 to -0.0004442	0.0004	Balanced logistic stacker
ID	Synthesis +20 labels	Dice	-0.0004188	-0.0009145 to 0.0000618	0.9248	Not significant
ID	Synthesis +20 labels	Brier	-0.0000254	-0.0000592 to 0.0000097	0.5607	Not significant
OOD	Synthesis	Dice	+0.0004969	-0.0004190 to 0.0013275	0.6487	Not significant
OOD	Synthesis	Brier	+0.0000266	-0.0000244 to 0.0000747	0.9248	Not significant
OOD	Balanced stacker	Dice	+0.0446051	0.0330824 to 0.0557834	0.0007	QGate
OOD	Balanced stacker	Brier	+0.0000635	-0.0003377 to 0.0004645	0.9799	Not significant
OOD	Synthesis +20 labels	Dice	-0.0008848	-0.0029309 to 0.0015060	0.9248	Not significant
OOD	Synthesis +20 labels	Brier	-0.0001685	-0.0002508 to 0.0000836	0.0242	Synthesis +20 labels

Positive effects favour QGate: QGate minus comparator for Dice and comparator minus QGate for Brier. Confidence intervals are unadjusted; inference uses Holm-adjusted p values.

Table 5: Original core macro outcomes

Domain	Method	Macro Dice (95% CI)	Macro Brier (95% CI)
ID	Synthesis	0.962133 (0.960080-0.964012)	0.00146126 (0.00141575-0.00150629)
ID	QGate-MRI-R	0.962602 (0.960724-0.964346)	0.00143570 (0.00139484-0.00147738)
ID	Balanced logistic stacker	0.965062 (0.962636-0.967132)	0.00095991 (0.00092036-0.00100571)
ID	Synthesis +20 labels	0.963021 (0.960939-0.964851)	0.00141031 (0.00135602-0.00146769)
OOD	Synthesis	0.822745 (0.813871-0.831997)	0.00732618 (0.00696937-0.00764199)
OOD	QGate-MRI-R	0.823242 (0.814726-0.832096)	0.00729959 (0.00695495-0.00761790)
OOD	Balanced logistic stacker	0.778637 (0.765841-0.792219)	0.00736314 (0.00689287-0.00782960)
OOD	Synthesis +20 labels	0.824127 (0.814575-0.833274)	0.00713105 (0.00683601-0.00739843)

Entries are seed means with unadjusted 95% bootstrap confidence intervals. Lower Brier is better.

Figure 2 displays the prespecified seed-level effects for the two locked Holm families.

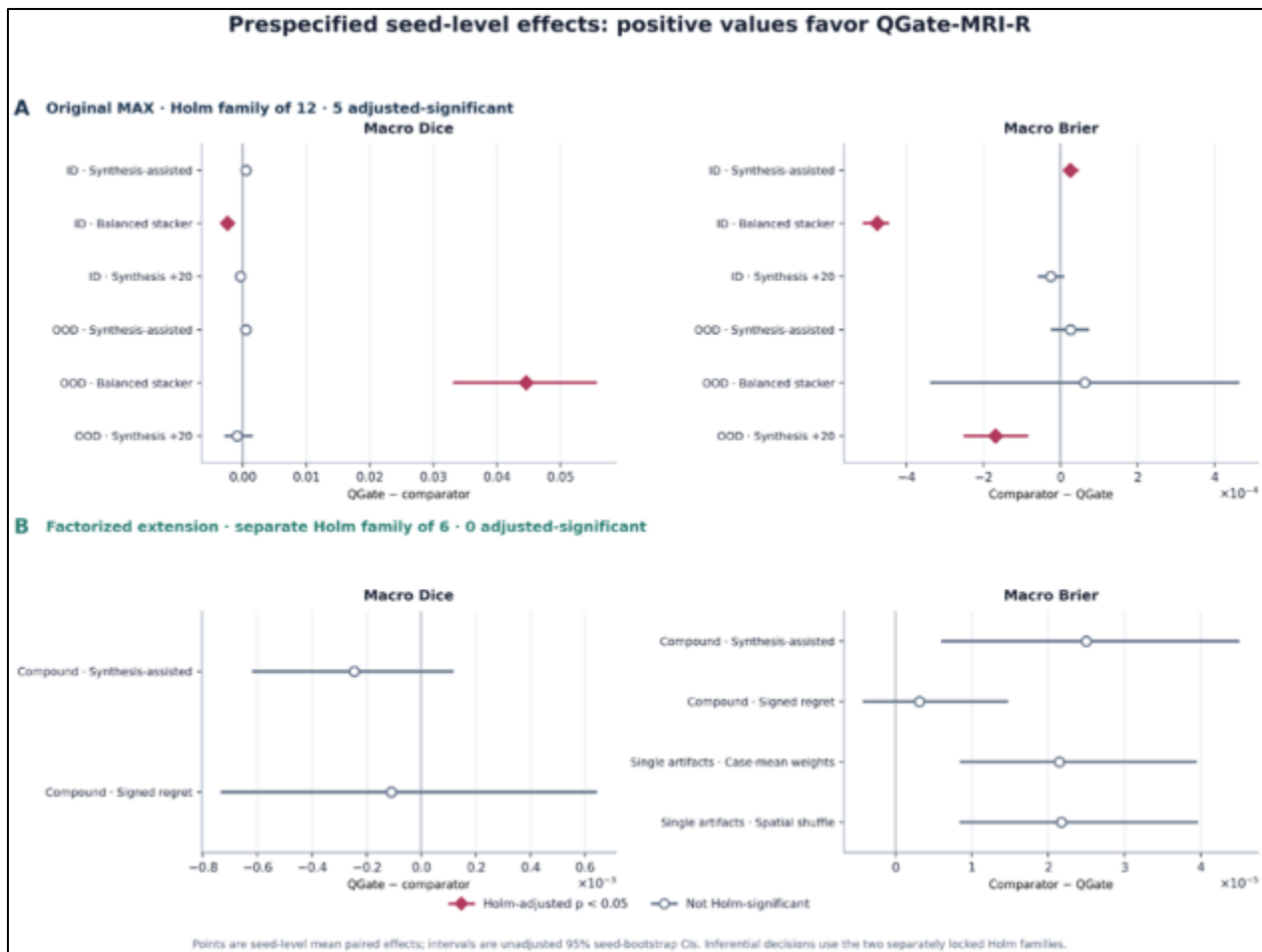


Figure 2: Prespecified seed-level effects in the two separately locked Holm families. Effects are oriented so positive values favour QGate. Diamonds indicate Holm-significant comparisons; confidence intervals are unadjusted 95% seed-bootstrap intervals. The families are displayed together for audit but were not pooled.

3.2 Router behaviour in Study 1

Table 6 reports regional router weights and benefit-ranking diagnostics.

Table 6: Original router weight and benefit-ranking diagnostics.

Region	ID mean w	ID within SD	OOD mean w	OOD within SD	ID AP	OOD AP
WT	0.6647	0.0128	0.6682	0.0251	0.741	0.654
TC	0.8170	0.0062	0.8187	0.0180	0.813	0.740
ET	0.8588	0.0047	0.8595	0.0093	0.788	0.742

AP denotes case-level average precision for the benefit indicator. Mean weight and within-map standard deviation refer to the calibrated synthesis weight.

Mean QGate synthesis weights for WT/TC/ET were 0.6647/0.8170/0.8588 in distribution and 0.6682/0.8187/0.8595 under generator B. Corresponding within-map standard deviations were only 0.0128/0.0062/0.0047 and 0.0251/0.0180/0.0093. Average precision was 0.741/0.813/0.788 in distribution and 0.654/0.740/0.742 under generator B; because positive-class prevalence was not reported, these values are interpreted descriptively. The small spatial variation suggested that calibrated weighting was already close to region-global.

Figure 3 summarises regional router diagnostics in distribution and under generator B.

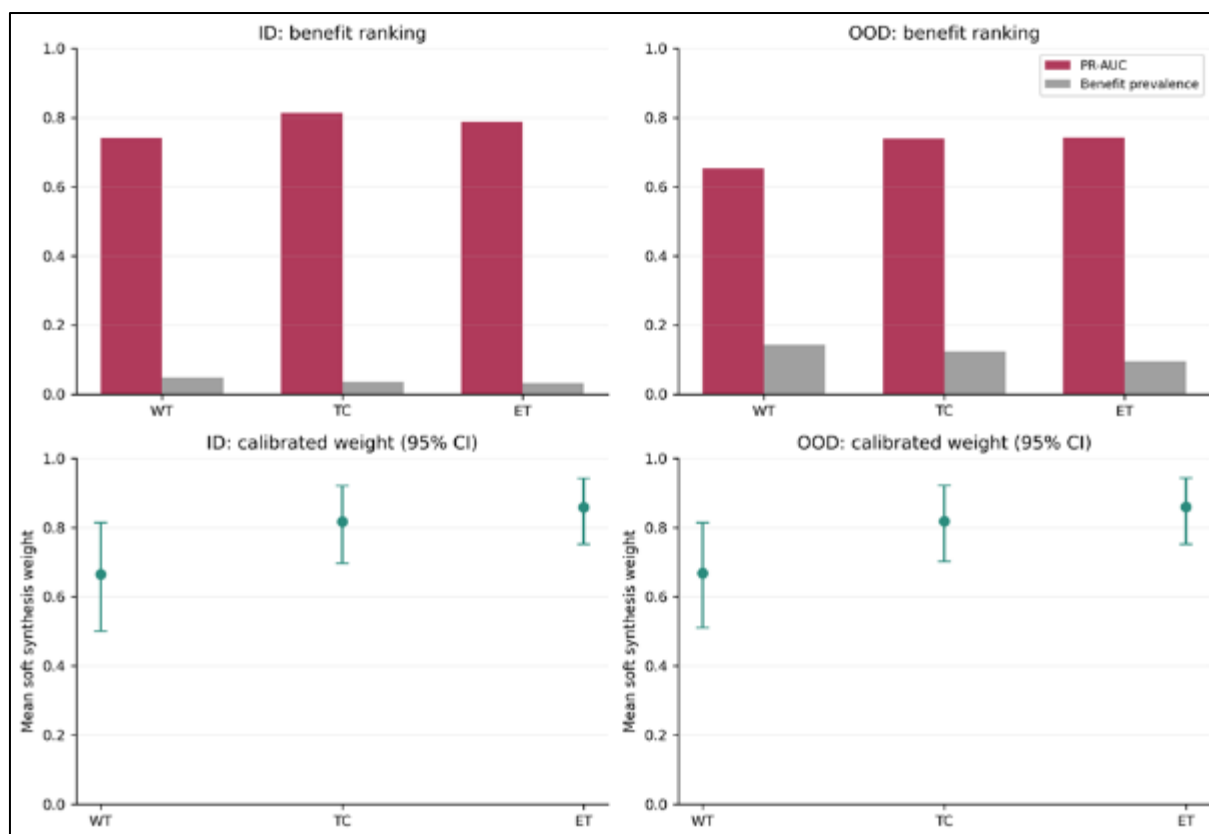


Figure 3: Original regional router diagnostics under in-distribution and generator-B shifted data. Scores, calibrated weights, and benefit labels have distinct meanings. Diagnostic discrimination alone does not establish that spatial placement improves the outcome.

3.3 Factorised falsification extension

3.3.1 Primary family

All 16 final pipeline runs and 8,320 paired case-condition evaluations were completed, and every prespecified exposure invariance check passed. None of the six extension comparisons was Holm-significant. The complete family is shown in Table 7; compound outcomes and condition-specific behaviour are reported in Table 8 and Figures 4 and 5.

Table 7: Complete extension six-test primary family.

Context	Comparator	Metric	Effect	95% CI	Raw p	Holm p
Compound	Synthesis	Dice	-0.00024492	-0.00061967 to 0.00011915	0.1297	0.2886
Compound	Synthesis	Brier	+0.00002501	0.00000597 to 0.00004509	0.0290	0.1160
Compound	Signed regret blend	Dice	-0.00010937	-0.00073530 to 0.00064462	0.0962	0.2886
Compound	Signed regret blend	Brier	+0.00000315	-0.00000427 to 0.00001479	0.4954	0.4954
Pooled artefacts	Case-mean QGate	Brier	+0.00002147	0.00000845 to 0.00003950	0.0092	0.0551
Pooled artefacts	Shuffled QGate	Brier	+0.00002174	0.00000844 to 0.00003967	0.0110	0.0551

Positive effects favour QGate. The two locality intervals excluded zero before multiplicity adjustment, but both Holm-adjusted p values were 0.0551.

Table 8: Compound available-input shift: selected descriptive outcomes.

<i>Method</i>	<i>Macro Dice (95% CI)</i>	<i>Macro Brier (95% CI)</i>
Synthesis	0.903082 (0.899757-0.906410)	0.00339356 (0.00330661-0.00347535)
QGate	0.902837 (0.899411-0.906173)	0.00336856 (0.00329296-0.00343870)
Case-mean QGate	0.902878 (0.899586-0.906190)	0.00337612 (0.00329563-0.00345153)
Shuffled QGate	0.902890 (0.899522-0.906164)	0.00337614 (0.00329601-0.00345198)
Signed regret blend	0.902947 (0.899553-0.906119)	0.00337170 (0.00329082-0.00344676)
Natural stacker	0.895603 (0.892156-0.898669)	0.00306985 (0.00298157-0.00317357)
Balanced stacker	0.896610 (0.893194-0.900018)	0.00307627 (0.00298618-0.00317306)
Synthesis +20	0.904261 (0.900866-0.907642)	0.00332994 (0.00325270-0.00339942)

Intervals are unadjusted 95% seed-bootstrap confidence intervals. Only the four prespecified QGate compound comparisons in Table 7 are inferential.

Under the compound shift, QGate macro Dice was 0.902837 versus 0.903082 for ungated synthesis-assisted segmentation, an effect of -0.00024492 (Holm-adjusted $p = 0.2886$). Macro Brier was 0.00336856 versus 0.00339356, an improvement of 0.00002501 that did not remain significant after adjustment (Holm-adjusted $p = 0.1160$). QGate was statistically indistinguishable from the signed logistic regret blend on both endpoints. Natural and balanced stackers achieved lower Brier but lower Dice, whereas synthesis with 20 additional labels had the highest displayed Dice. Rankings therefore depended on both condition and endpoint.

Under FLAIR dropout, all 15 methods had a macro Dice of zero. This shared simulator-specific collapse requires diagnostic investigation of channel scaling, class probabilities, and prediction frequencies; the dropout result is therefore not used for comparative inference.

Figure 4 compares selected methods across the factorised acquisition conditions.

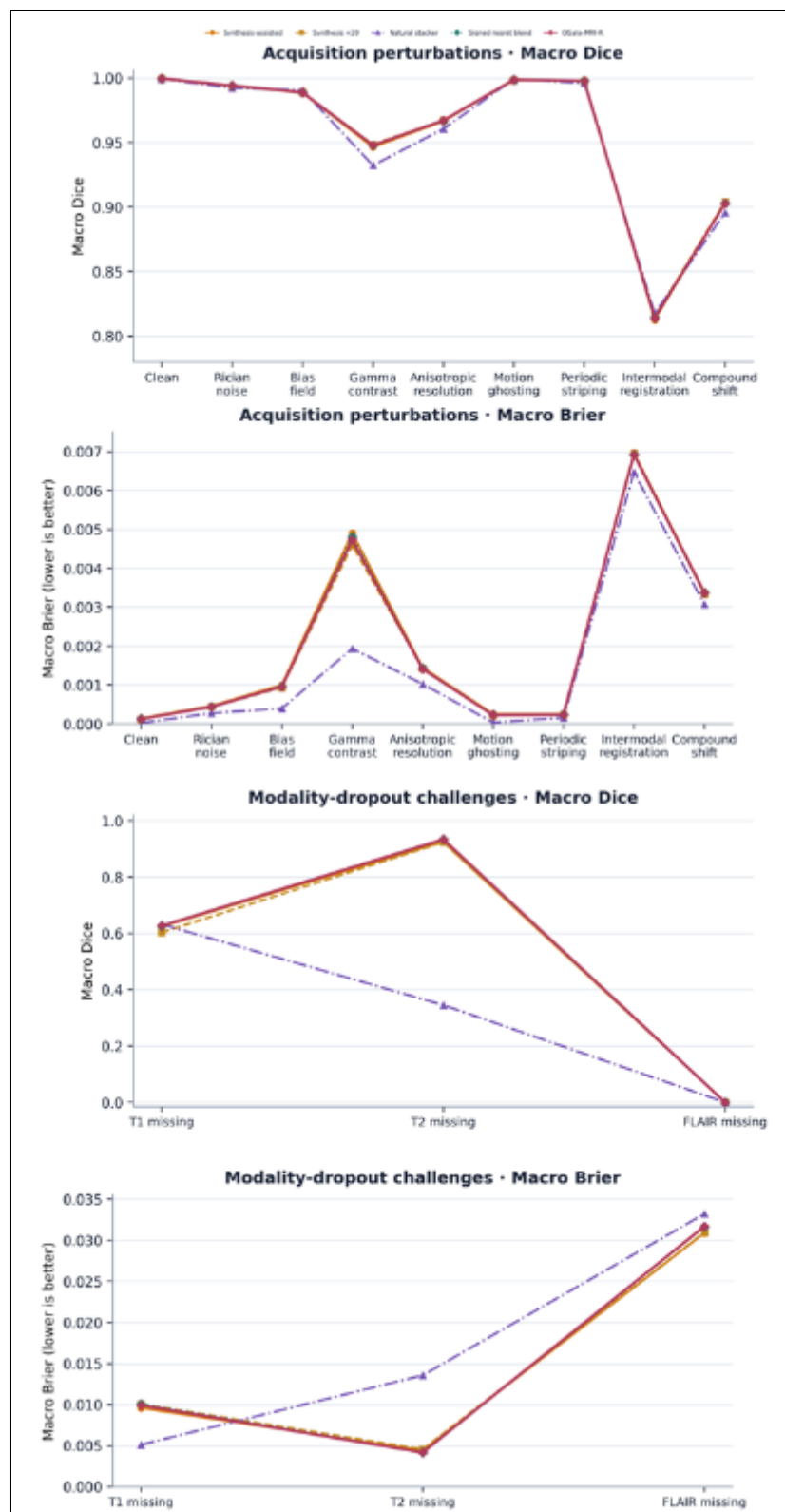


Figure 4: Factorised-condition performance across methods. The one-factor acquisition stresses reveal heterogeneous failure modes but do not decompose the original generator-B shift. Dropouts and descriptive secondary comparisons were not added to the primary family.

Figure 5 shows the prespecified qualitative audit for the fixed seed and case.

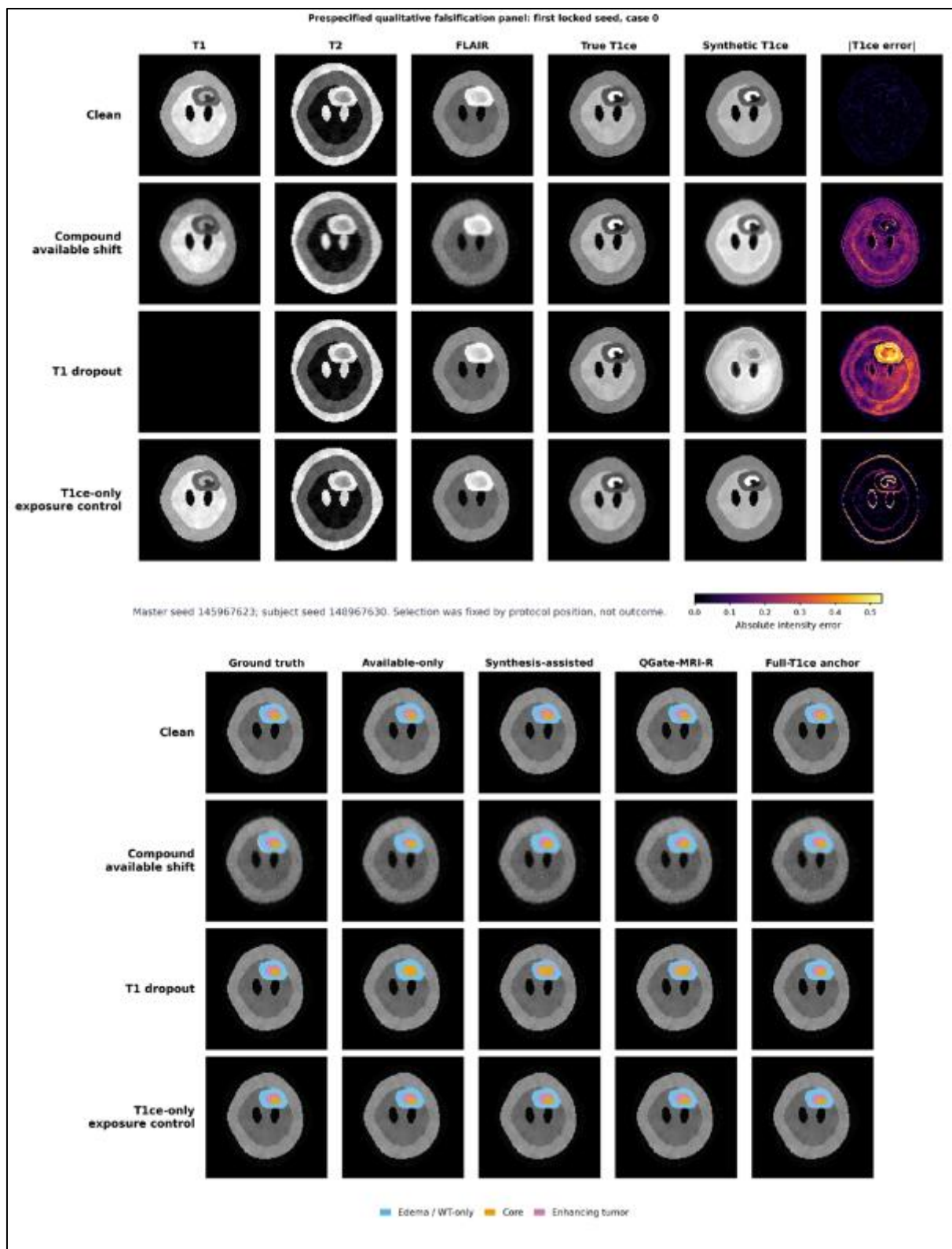


Figure 5: Prespecified qualitative reconstruction and segmentation audit for final seed 145967623, test case 0. The seed, case, and conditions were fixed before outcome inspection. The panel illustrates clean, compound, T1-dropout, and T1ce-only exposure inputs and is not evidence of population performance.

3.3.2 Locality analysis

Across the seven pooled single-artefact conditions, native QGate weights reduced macro Brier relative to case-mean weights by 0.00002147 (unadjusted 95% CI 0.00000845 to 0.00003950; raw $p = 0.0092$; Holm-adjusted $p = 0.0551$) and relative to histogram-preserving shuffled weights by 0.00002174 (0.00000844 to 0.00003967; raw $p = 0.0110$;

Holm-adjusted $p = 0.0551$). The effect directions and unadjusted intervals were compatible with a small local contribution, but repeatable locality was not established under the completed multiplicity family.

Table 9 separates across-seed variation from within-map spatial variation.

Table 9: Mean synthesis weights: across-seed and within-map variation.

<i>Condition</i>	<i>Region</i>	<i>Mean weight</i>	<i>Between-seed SD</i>	<i>Mean within-map SD</i>
Clean	WT	0.5965	0.3972	0.0050
Clean	TC	0.9133	0.1823	0.0010
Clean	ET	0.9073	0.1555	0.0009
Compound	WT	0.5962	0.3972	0.0099
Compound	TC	0.9134	0.1821	0.0050
Compound	ET	0.9075	0.1551	0.0058
T1 dropout	WT	0.6124	0.4024	0.0340
T1 dropout	TC	0.9148	0.1798	0.0132
T1 dropout	ET	0.9083	0.1533	0.0093

Each row first averages case-level diagnostics within a complete training seed. Between-seed standard deviation combines fitted-model, regenerated case-composition, and generator variability; within-map standard deviation describes spatial variation inside a case.

On clean cases, mean WT/TC/ET weights were 0.5965/0.9133/0.9073, with between-seed standard deviations of 0.3972/0.1823/0.1555 but mean within-map standard deviations of only 0.0050/0.0010/0.0009. Compound within-map values remained 0.0099/0.0050/0.0058, and even T1 dropout increased them only to 0.0340/0.0132/0.0093. Across-seed variation in mean regional weights was therefore much larger than within-map spatial variation; because each seed regenerated fitted models and evaluation cases, this contrast does not isolate model-fitting instability from case-composition or generator variability.

Figure 6 compares native QGate weights with the two locality-removal controls.

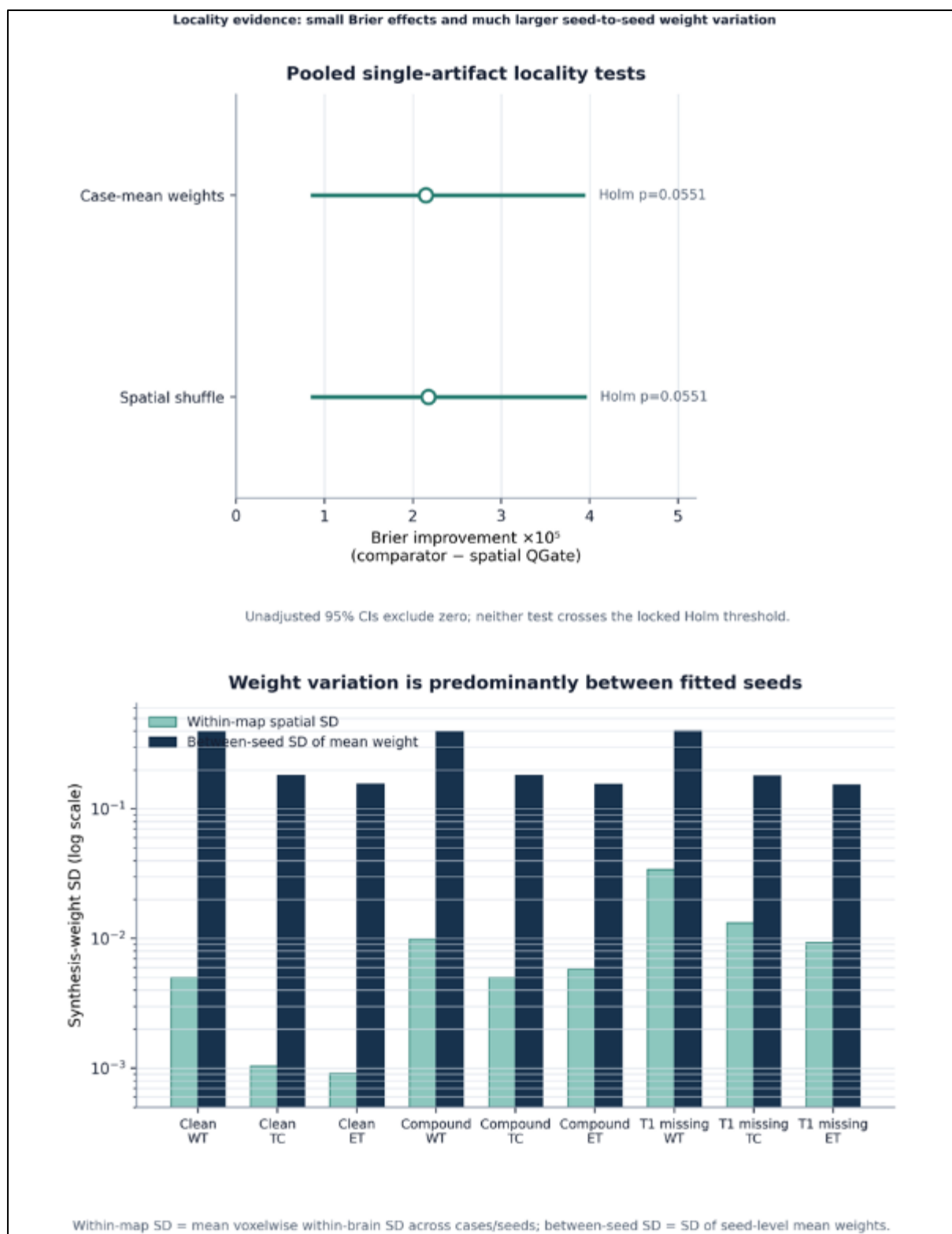


Figure 6: Locality audit using native, case-mean, and histogram-preserving shuffled QGate weights. The controls preserve the amount of synthesis weighting while removing local placement. The two pooled Brier comparisons did not meet the Holm-corrected threshold.

Figure 7 shows that the fixed-case weight maps had narrow dynamic ranges, consistent with the aggregate within-map diagnostics.

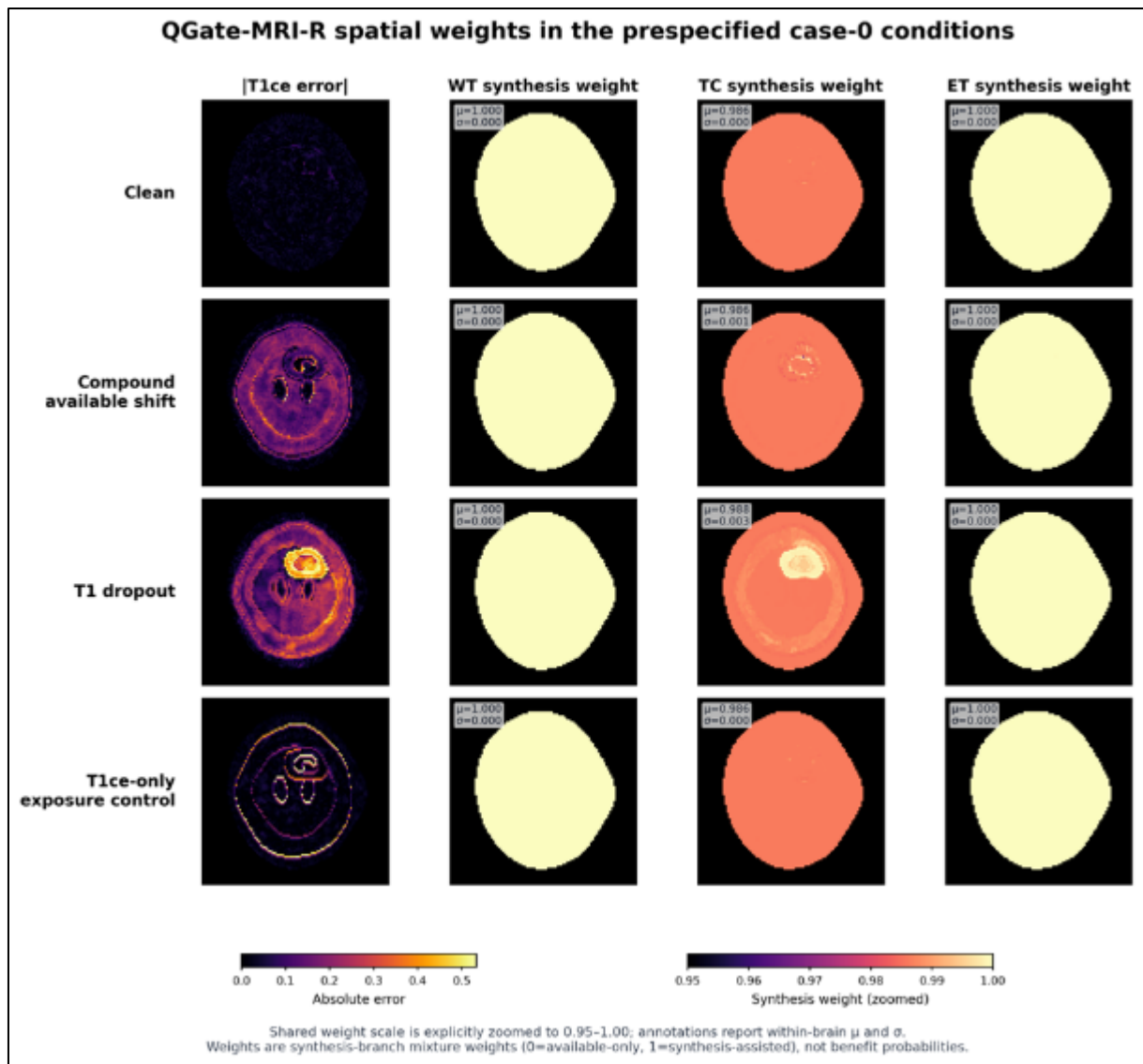


Figure 7: Prespecified QGate weight maps for final seed 145967623, test case 0. Maps are shown for the same fixed case across selected conditions. Their narrow dynamic ranges are consistent with the aggregate within-map diagnostics.

3.3.3 Stronger stacker controls

Natural-prevalence and signed-regret logistic constructions separated class balancing, feature signing, and calibration restriction. On the compound shift, natural and balanced direct stackers achieved lower Brier than QGate but lower Dice, while the signed regret blend tracked QGate closely. Across individual conditions, rankings changed. The evidence therefore concerns endpoint- and shift-specific trade-offs rather than an intrinsic advantage of tree routing over logistic meta-models.

Figure 8 compares QGate with the stacker variants across factorised conditions.



Figure 8: QGate and stacker variants across factorised conditions. The figure is descriptive outside the two prespecified QGate-versus-signed-regret compound tests. Lower Brier can coexist with lower thresholded Dice.

3.4 Exposure and identifiability controls

3.4.1 T1ce-only exposure

The exposure control changed only the hidden true T1ce target. Observed T1, T2, and FLAIR were bitwise identical to clean, so the synthesiser input, synthetic image, available-only prediction, synthesis-assisted prediction, QGate weights and output, and stacker outputs were forced to remain unchanged. Reconstruction fidelity nevertheless decreased because the unchanged synthetic image was compared with a translated target. The observed-T1ce anchor changed because it alone consumed the perturbed channel.

Table 10 reports the exact target-only exposure audit.

Table 10: T1ce-only exposure control.

<i>Quantity</i>	<i>Clean</i>	<i>Exposed</i>	<i>Change / audit</i>
Synthesis PSNR (dB)	29.229	17.007	-12.222
Synthesis SSIM	0.908679	0.758501	-0.150177
Observed-T1ce reference macro Dice	0.999519	0.865601	-0.133918
Observed-T1ce reference macro Brier	0.00011142	0.00274392	+0.00263249
True-T1ce MAE from clean	0 by definition	0.058447	Target changed
Deployable invariance checks	640/640 expected	640/640 passed	640/640 passed

Change is exposed minus clean. The 640 entries are case-level invariance checks across 40 cases and 16 final seeds.

PSNR decreased by 12.22 dB and SSIM by 0.150 while all 640 deployable invariance checks passed. This constructed control shows that a reconstruction metric can change even when there is no computational path from the altered target to the deployed prediction.

Figure 9 summarises the target-only exposure control.

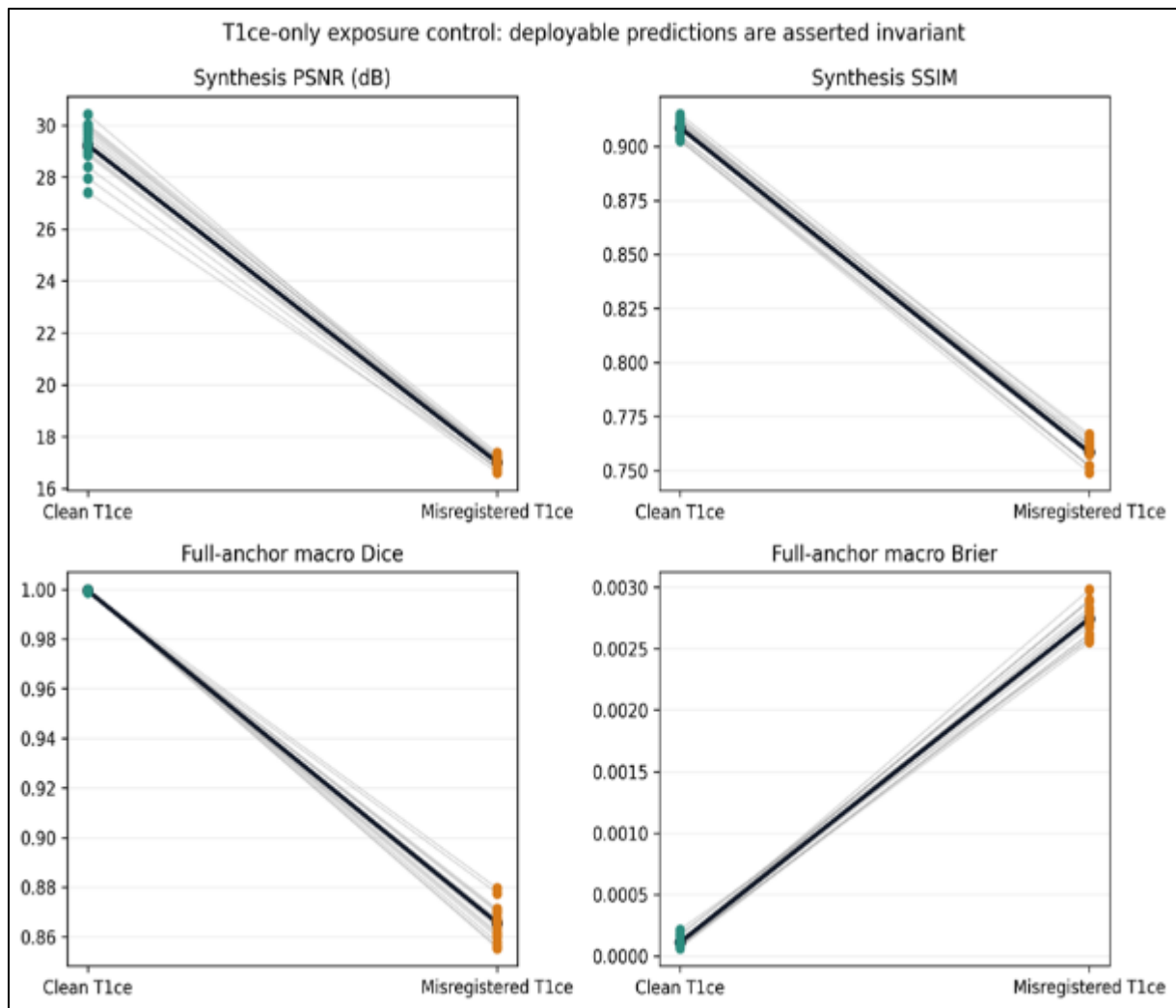


Figure 9: T1ce-only exposure control separating reconstruction fidelity from deployable segmentation utility. Only the hidden true T1ce and the observed-T1ce anchor changed. All asserted missing-T1ce prediction paths remained bitwise invariant for every case and seed.

3.4.2 Identical-observable ET ambiguity

All 32 counterfactual pairs passed the exact identity, equal-volume, and disjointness checks. Available modalities matched bitwise with maximum absolute difference zero, WT and TC were identical, and each hidden ET alternative contained 54–99 pixels. For a common probabilistic prediction, the ambiguity-support Brier lower bound was 0.25; after normalisation over each brain, the whole-brain ET bound ranged from 0.006239 to 0.011480. A shared binary hard prediction has a corresponding ambiguity-support Brier minimum of 0.50. The maximum paired-mean and worst-member ET Dice for a shared deterministic prediction was $2/3$.

Table 11 defines the deterministic method families covered by the collision bounds.

Table 11: Applicability of the exact counterfactual bounds.

Method family	Inference mapping	True T1ce observed?	Construction-specific bound
Available-only	$f(X)$	No	ET Brier on ambiguity support ≥ 0.25 ; paired ET Dice $\leq 2/3$
Synthesis-assisted	$g(X, \text{synthesis}(X))$	No	ET Brier on ambiguity support ≥ 0.25 ; paired ET Dice $\leq 2/3$
QGate-MRI-R	$\text{router}(X, \text{synthesis}(X), \text{branch outputs})$	No	ET Brier on ambiguity support ≥ 0.25 ; paired ET Dice $\leq 2/3$

Logistic stacking	stacker(branch derived from X)	outputs	No	ET Brier on ambiguity support ≥ 0.25 ; paired ET Dice $\leq 2/3$
Full-modality reference	$h(X, \text{true T1ce})$		Yes	Not applicable

The bounds apply to deterministic implementations, or fixed predictive distributions, whose inference uses only the identical observed inputs and no hidden channel or distinguishing metadata.

Figure 10 shows the first prespecified identical-observable pair and its hidden alternatives.

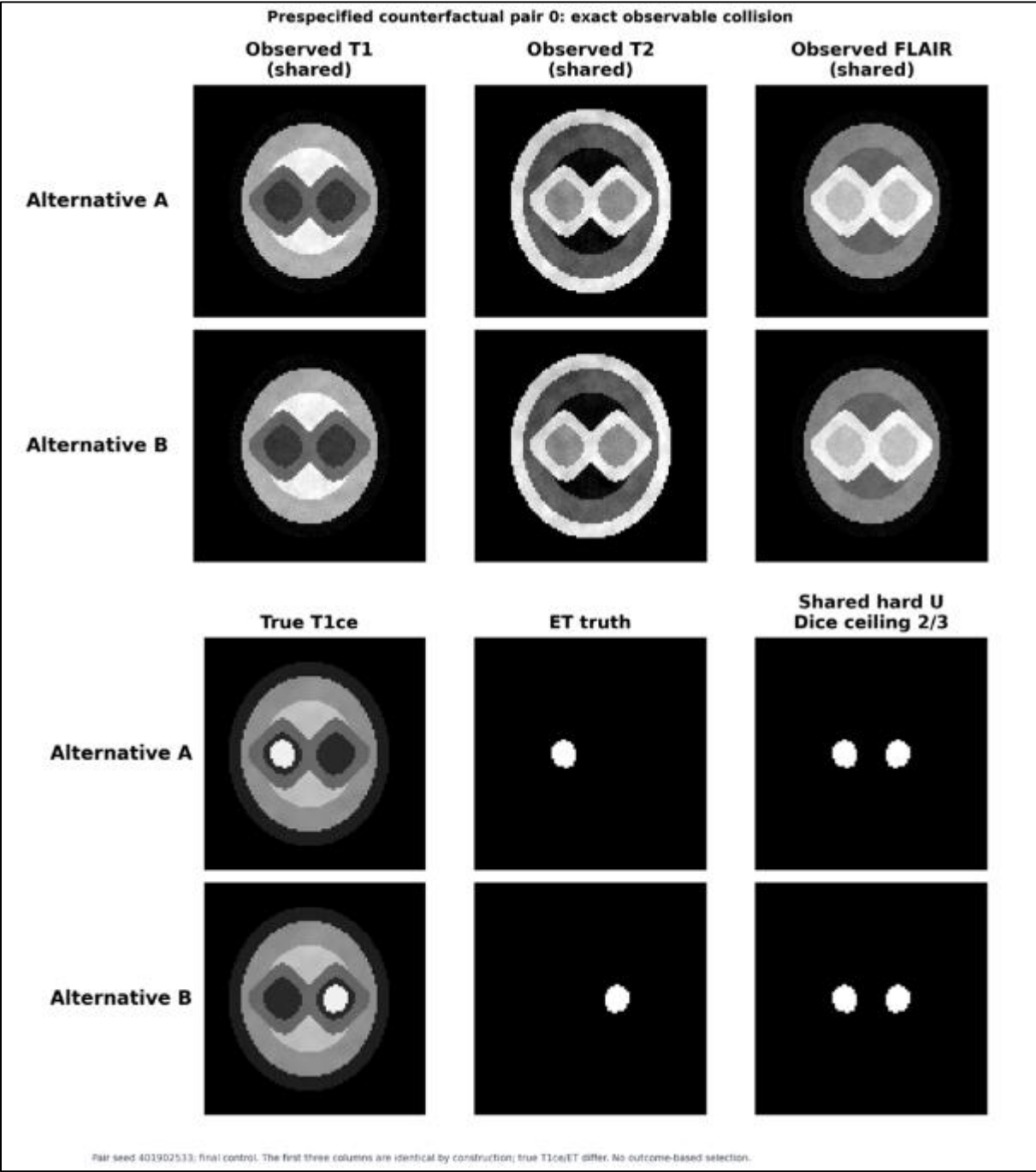


Figure 10: Exact identical-observable counterfactual pair and hidden ET alternatives. Observed T1, T2, and FLAIR and the WT/TC masks are identical, while the equal-volume hidden ET/T1ce supports are disjoint. The visual is the first prespecified pair from the 32-pair final run.

3.5 Claim boundaries supported by the evidence

Table 12 summarises the claim boundaries established by the complete evidence. These decisions are restricted to the procedural generators, fitted models, label budgets, and locked controls used here; they do not extend to clinical MRI.

Table 12: Evidence-supported claim boundaries.

Claim	Decision	Controlling evidence
QGate broadly outperforms synthesis	Rejected	Mixed original family; extension compound Dice/Brier not Holm-significant
QGate improves original ID probabilistic accuracy	Narrowly supported	ID Brier versus ungated synthesis significant; label-matched +20 comparison not significant
QGate robustly localises synthesis benefit	Not established	Both locality tests Holm-adjusted $p = 0.0551$; within-map SD small; across-seed means variable
Direct stacking is generally inferior	Rejected	Best original ID outcomes; lower compound Brier in extension; Dice failures were condition-specific
Reconstruction fidelity determines utility	Not sufficient under target-only control	Fidelity decreased while all deployable predictions remained bitwise invariant
Routing can recover unobserved ET without evidence in X	Impossible under the defined collision	ET Brier on ambiguity support ≥ 0.25 and paired ET Dice $\leq 2/3$
Method is clinically ready	Unsupported	No patients, 3D networks, sites, scanners, prospective data, or clinical outcomes

Here, 'unsupported' and 'not established' denote outcomes of completed locked analyses, not missing analyses.

4 DISCUSSION

The principal finding is that the separately locked extension did not establish the proposed spatial-routing mechanism. None of its six prespecified tests remained significant after Holm correction. Under the compound shift, the small numerical Brier reduction did not establish corrected superiority, and Dice did not favour QGate. Native maps showed small unadjusted Brier advantages over case-mean and shuffled counterparts, but both locality comparisons narrowly missed the familywise threshold. These results do not prove that local structure has no effect, but repeatable voxelwise benefit was not established by the completed experiment.

The extension does not invalidate the original in-distribution Brier result. That improvement over ungated synthesis-assisted segmentation remains supported within the original 12-test family, but its interpretation is narrower than a spatial-locality claim. QGate accessed 20 second-level labels beyond the frozen 48-label synthesis branch, and the label-matched synthesis +20 comparison was not significant in distribution. A calibrated regional blend can therefore improve probabilistic loss without providing a valid map of where a synthetic image is useful.

The comparison with stacking also separates performance from mechanism. The balanced logistic stacker performed best in distribution but had markedly lower Dice under generator B, while its shifted Brier was statistically indistinguishable from QGate. That Dice difference is relevant for the tested shift, yet it does not show that a restricted tree router is generally superior to stacked generalisation. In the extension, natural and balanced stackers produced the lowest compound Brier among supervised fusion methods shown, while synthesis with additional labels produced the highest displayed Dice. Probability quality and thresholded overlap therefore answer different questions.

The learned weight structure reinforces this interpretation. Seed-level aggregate weights varied much more than weights within individual maps. Because each seed regenerated training partitions and evaluation cases, the study cannot attribute this difference solely to fitted-model instability. Combined with the case-mean and shuffled controls, the most defensible description in this simulator is a predominantly region-global blend with, at most, weakly suggestive local refinement.

The exposure control clarifies why reconstruction metrics should not be treated as automatic surrogates for downstream utility. A target-only perturbation reduced PSNR by 12.22 dB while every deployed missing-T1ce input and prediction remained identical. Under this constructed perturbation, reconstruction fidelity was therefore not a sufficient surrogate for deployable utility. Claims linking PSNR or SSIM to segmentation should identify which variables changed and demonstrate downstream effects in a paired design.

The identical-observable construction supplies a stronger information boundary. When two hidden ET states produce the same observed modalities, every deterministic synthesiser, router, stacker, or segmenter restricted to those modalities must return the same prediction. Greater complexity can change priors or average behaviour but cannot

recover sample-specific evidence absent from the input. The 0.25 ambiguity-support Brier lower bound and 2/3 paired Dice ceiling are statements about the defined equal-prior collision, not universal limits for clinical MRI.

These findings complement missing-modality and uncertainty research by shifting emphasis from whether a router changes a benchmark score to whether an experiment isolates the claimed mechanism [3, 16, 17, 23]. Related learning-to-defer formulations likewise formalise routing among multiple experts [38]. A robust evaluation should lock an unseen shift, use complete training pipelines as the inferential unit, equalise labelled-data access, include flexible and restricted second-level comparators, remove spatial structure while preserving weight distributions, and state the actual information path. Procedural phantoms are useful for these operations, but they remain a preliminary mechanism test rather than patient validation.

A deployment choice also cannot be inferred from one headline metric. In the original experiment, a user prioritising in-distribution Dice or Brier would select stacking; one prioritising generator-B Dice would avoid that stacker; and one prioritising shifted Brier would select synthesis with additional labels. Under the factorised compound shift, stackers had the lowest displayed Brier while additional-label synthesis had the highest Dice. These simulator-specific trade-offs are valuable for designing later validation, not for prescribing clinical use.

Limitations

This study used two-dimensional procedural phantoms with simplified anatomy, pathology, labels, and acquisition physics. It included no patients, clinical sites, scanners, reader studies, prospective workflow, or clinical outcomes; consequently, no inference about clinical performance, safety, or patient benefit is warranted. Tree-based synthesis and segmentation were selected for transparent CPU reproducibility rather than contemporary three-dimensional accuracy, and rankings may change with neural architectures or patient-trained representations. Artefact magnitudes and the compound condition were authored rather than sampled from real site, scanner, protocol, or demographic distributions. Pairing each condition to the same anatomy strengthens attribution within the simulator but is not external validation.

The experimental scope was restricted to missing T1ce, one WT/TC/ET hierarchy, fixed meta-training and calibration sizes, and deterministic inference. Other missingness patterns, label structures, budgets, and stochastic rules require separate evaluation. Study 2 was designed after Study 1 but locked separately before its 16 final seeds; the studies must not be pooled. The locality effects had unadjusted intervals excluding zero but Holm-adjusted p values of 0.0551. The implemented shuffle did not preserve spatial smoothness, and repeated or blockwise permutations were not evaluated. More seeds might change precision, yet the prespecified completed result remains non-significant after multiplicity correction. No prospective power analysis, equivalence margin, or multiplicity-adjusted confidence intervals were reported; the unadjusted intervals are therefore descriptive and non-significance is not evidence of equivalence.

The target-only exposure experiment demonstrates metric-utility decoupling for a deliberately constructed perturbation and does not estimate its clinical frequency. Likewise, the counterfactual theorem proves a bound under exact observable identity but does not estimate collision prevalence in patient data. The test-label selector is an unattainable choice between two branches only and is not a universal ceiling, while the observed-T1ce model is a full-information reference rather than a deployable missing-modality method.

Reporting transparency

Table 13 records reporting items based on the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) and the corresponding applicability boundaries.

Table 13: CLAIM-oriented reporting applicability for this non-clinical simulation study.

Reporting item	Status	Implementation / boundary
Clinical intended use	Not claimed	Procedural mechanism benchmark only
Data source / eligibility	Not applicable to patients	Procedural generators only; no patient eligibility criteria
Reference standard	Exact simulator labels	WT/TC/ET generated with each phantom
Train/validation/test separation	Reported	48/48/12/8 fitting partitions; test disjoint
External validation	Not performed	Generator B and artefacts are authored shifts, not external cohorts

Sample size	Reported	16 seeds per study; no pooling; 32 counterfactual pairs
Performance measures	Partly reported	Primary macro Dice/Brier shown; complete prespecified outcomes require a supplement
Uncertainty / multiplicity	Reported	Seed bootstrap CIs; Wilcoxon; separate Holm families
Failure analysis	Partly investigated	FLAIR collapse, locality controls, exposure, and collisions; collapse requires diagnosis
Code and reproducibility	Package described	Repository URL/DOI and version identifier are not reported

Items requiring patient enrolment, clinical sites, scanners, reader studies, or workflow evaluation are reported as not applicable or not performed rather than implied by phantom results [37].

5 CONCLUSION

Across two separately locked 16-seed phantom studies and a 32-pair analytic control, QGate did not demonstrate robust spatial routing of synthesised T1ce information. Its original in-distribution Brier advantage over ungated synthesis-assisted segmentation was small and did not establish a voxelwise mechanism; none of six extension comparisons survived Holm correction. Across-seed variation in mean weights exceeded within-map spatial variation, but the design did not isolate fitted-model variability from regenerated case composition. Target-only exposure and identical-observable controls further showed that reconstruction fidelity need not track deployable utility and that unobserved evidence imposes construction-specific limits. The main contribution is a transparent framework for falsifying routing claims before patient-level validation, not evidence of clinical readiness or general QGate superiority.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The author declares no conflict of interest.

Data and code availability

The reproducibility package contains both locked protocols, source scripts, seed-specific configurations and manifests, per-case outputs, aggregate tables, calibration summaries, figures, and the manuscript builder.

Author contributions

Zahraa S. Abd-Al Ameer conceived the study, developed the methodology, performed the analyses, interpreted the results, and wrote and approved the manuscript.

Generative artificial intelligence

OpenAI ChatGPT was used for language editing and manuscript-formatting support. All numerical results were generated by the locked code and verified by the author, who takes full responsibility for the manuscript.

REFERENCES

- [1] Menze, B.H. et al. (2015) 'The multimodal brain tumor image segmentation benchmark (BRATS)', IEEE Transactions on Medical Imaging, Vol. 34, No. 10, pp. 1993–2024, doi:10.1109/TMI.2014.2377694.
- [2] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K. and Davatzikos, C. (2017) 'Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features', Scientific Data, Vol. 4, Article 170117, doi:10.1038/sdata.2017.117.
- [3] Havaei, M., Guizard, N., Chapados, N. and Bengio, Y. (2016) 'HeMIS: hetero-modal image segmentation', Medical Image Computing and Computer-Assisted Intervention, pp. 469–477, doi:10.1007/978-3-319-46723-8_54.
- [4] Dorent, R., Joutard, S., Modat, M., Ourselin, S. and Vercauteren, T. (2019) 'Hetero-modal variational encoder-decoder for joint modality completion and segmentation', Medical Image Computing and Computer Assisted Intervention, pp. 74–82, doi:10.1007/978-3-030-32245-8_9.
- [5] Chen, C., Dou, Q., Jin, Y., Liu, Q. and Heng, P.-A. (2022) 'Learning with privileged multimodal knowledge for unimodal segmentation', IEEE Transactions on Medical Imaging, Vol. 41, No. 3, pp. 621–632, doi:10.1109/TMI.2021.3119385.

- [6] Ding, Y., Yu, X. and Yang, Y. (2021) 'RFNet: region-aware fusion network for incomplete multi-modal brain tumor segmentation', *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3975–3984, doi:10.1109/ICCV48922.2021.00394.
- [7] Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z. and Zheng, Y. (2022) 'mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation', *Medical Image Computing and Computer Assisted Intervention*, pp. 107–117, doi:10.1007/978-3-031-16443-9_11.
- [8] Liu, H., Wei, D., Lu, D., Sun, J., Wang, L. and Zheng, Y. (2023) 'M3AE: multimodal representation learning for brain tumor segmentation with missing modalities', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 2, pp. 1657–1665, doi:10.1609/aaai.v37i2.25253.
- [9] Shi, J., Yu, L., Cheng, Q., Yang, X., Cheng, K.-T. and Yan, Z. (2024) 'M2FTrans: modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation', *IEEE Journal of Biomedical and Health Informatics*, Vol. 28, No. 1, pp. 379–390, doi:10.1109/JBHI.2023.3326151.
- [10] Zhou, T., Ruan, S. and Canu, S. (2019) 'A review: deep learning for medical image segmentation using multi-modality fusion', *Array*, Vol. 3–4, Article 100004, doi:10.1016/j.array.2019.100004.
- [11] Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E. and Merhof, D. (2024) 'Medical image segmentation review: the success of U-Net', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 12, pp. 10076–10095, doi:10.1109/TPAMI.2024.3435571.
- [12] Chatsias, A., Joyce, T., Giuffrida, M.V. and Tsaftaris, S.A. (2018) 'Multimodal MR synthesis via modality-invariant latent representation', *IEEE Transactions on Medical Imaging*, Vol. 37, No. 3, pp. 803–814, doi:10.1109/TMI.2017.2764326.
- [13] Huang, P., Li, D., Jiao, Z., Wei, D., Cao, B., Mo, Z., Wang, Q., Zhang, H. and Shen, D. (2022) 'Common feature learning for brain tumor MRI synthesis by context-aware generative adversarial network', *Medical Image Analysis*, Vol. 79, Article 102472, doi:10.1016/j.media.2022.102472.
- [14] Zhou, T., Fu, H., Chen, G., Shen, J. and Shao, L. (2020) 'Hi-Net: hybrid-fusion network for multi-modal MR image synthesis', *IEEE Transactions on Medical Imaging*, Vol. 39, No. 9, pp. 2772–2781, doi:10.1109/TMI.2020.2975344.
- [15] Li, H.B., Conte, G.M., Hu, Q. et al. (2024) 'The Brain Tumor Segmentation (BraTS) Challenge 2023: brain MR image synthesis for tumor segmentation (BraSyn)', arXiv:2305.09011.
- [16] Mehta, R., Christinck, T., Nair, T., Bussy, A., Premasiri, S., Costantino, M., Chakravarthy, M.M., Arnold, D.L., Gal, Y. and Arbel, T. (2022) 'Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference', *IEEE Transactions on Medical Imaging*, Vol. 41, No. 2, pp. 360–373, doi:10.1109/TMI.2021.3114097.
- [17] Liu, W., Zhang, Z., Deng, F. and Xiao, Y. (2026) 'Uncertainty-aware feature mapping and adaptive inference for brain tumor segmentation with missing contrast-enhanced T1-weighted imaging', *Frontiers in Neuroscience*, Vol. 20, Article 1828237, doi:10.3389/fnins.2026.1828237.
- [18] Zhou, T. and Zhu, S. (2023) 'Uncertainty quantification and attention-aware fusion guided multi-modal MR brain tumor segmentation', *Computers in Biology and Medicine*, Vol. 163, Article 107142, doi:10.1016/j.compbimed.2023.107142.
- [19] Zhu, Y., Li, K., Yu, L. and Heng, P.-A. (2026) 'Tackling missing modalities with memory-efficient modality-complementary prompt learning for robust brain tumor segmentation', *Expert Systems with Applications*, Vol. 308, Article 130976, doi:10.1016/j.eswa.2025.130976.
- [20] Almansour, A.G.M. and Alshomrani, F. (2026) 'Dynamic expert routing with boundary-aware decoding for accurate brain tumor segmentation from incomplete MRI data', *Scientific Reports*, Vol. 16, Article 16921, doi:10.1038/s41598-026-48187-x.
- [21] Zhou, T., Ruan, S., Meng, Y., Duan, J., Hu, H., Jiang, B., Ji, Z., Liu, B., Luo, T. and Lei, B. (2026) 'UTriGate-Net: uncertainty-aware brain tumor segmentation via triaxial context encoding and gated modality fusion', *Expert Systems with Applications*, Vol. 321, Article 132258, doi:10.1016/j.eswa.2026.132258.
- [22] Ali, D., Mian, A., Akhtar, N. and Hassan, G.M. (2026) 'D3Seg: dependency-aware diffusion for brain tumor segmentation with missing modalities', arXiv:2605.22249.
- [23] Wolpert, D.H. (1992) 'Stacked generalization', *Neural Networks*, Vol. 5, No. 2, pp. 241–259, doi:10.1016/S0893-6080(05)80023-1.
- [24] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) 'On calibration of modern neural networks', *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, pp. 1321–1330.

- [25] [25] Mehrtash, A., Wells, W.M., Tempny, C.M., Abolmaesumi, P. and Kapur, T. (2020) 'Confidence calibration and predictive uncertainty estimation for deep medical image segmentation', *IEEE Transactions on Medical Imaging*, Vol. 39, No. 12, pp. 3868–3878, doi:10.1109/TMI.2020.3006437.
- [26] Ding, Y., Liu, J., Xiong, J. and Shi, Y. (2020) 'Revisiting the evaluation of uncertainty estimation and its application to explore model complexity–uncertainty trade-off', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 22–31, doi:10.1109/CVPRW50498.2020.00010.
- [27] Collins, D.L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J. and Evans, A.C. (1998) 'Design and construction of a realistic digital brain phantom', *IEEE Transactions on Medical Imaging*, Vol. 17, No. 3, pp. 463–468, doi:10.1109/42.712135.
- [28] Kwan, R.K.-S., Evans, A.C. and Pike, G.B. (1999) 'MRI simulation-based evaluation of image-processing and classification methods', *IEEE Transactions on Medical Imaging*, Vol. 18, No. 11, pp. 1085–1097, doi:10.1109/42.816072.
- [29] Gudbjartsson, H. and Patz, S. (1995) 'The Rician distribution of noisy MRI data', *Magnetic Resonance in Medicine*, Vol. 34, No. 6, pp. 910–914, doi:10.1002/mrm.1910340618.
- [30] Sled, J.G., Zijdenbos, A.P. and Evans, A.C. (1998) 'A nonparametric method for automatic correction of intensity nonuniformity in MRI data', *IEEE Transactions on Medical Imaging*, Vol. 17, No. 1, pp. 87–97, doi:10.1109/42.668698.
- [31] Geurts, P., Ernst, D. and Wehenkel, L. (2006) 'Extremely randomized trees', *Machine Learning*, Vol. 63, pp. 3–42, doi:10.1007/s10994-006-6226-1.
- [32] Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) 'Image quality assessment: from error visibility to structural similarity', *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp. 600–612, doi:10.1109/TIP.2003.819861.
- [33] Taha, A.A. and Hanbury, A. (2015) 'Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool', *BMC Medical Imaging*, Vol. 15, Article 29, doi:10.1186/s12880-015-0068-x.
- [34] Wilcoxon, F. (1945) 'Individual comparisons by ranking methods', *Biometrics Bulletin*, Vol. 1, No. 6, pp. 80–83.
- [35] Holm, S. (1979) 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics*, Vol. 6, No. 2, pp. 65–70.
- [36] Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York.
- [37] Mongan, J., Moy, L. and Kahn, C.E. Jr (2020) 'Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers', *Radiology: Artificial Intelligence*, Vol. 2, No. 2, Article e200029, doi:10.1148/ryai.2020200029.
- [38] Verma, R., Barrejón, D. and Nalisnick, E. (2023) 'Learning to defer to multiple experts: consistent surrogate losses, confidence calibration, and conformal ensembles', *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, PMLR, Vol. 206, pp. 11415–11434.