Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Risk-Screened Multimodal Attribution for Glioma-like Segmentation: A Phantom-Based Study of Explanation Stability

zifan guo

2277135201@qq.com

University College London

piao ti

University College London

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11891847/v1

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

Additional Declarations: The authors declare no competing interests.

Risk-Screened Multimodal Attribution for Glioma-like Segmentation: A Phantom-Based Study of Explanation Stability

Zifan Guo* Piao Ti

University College London (UCL), London, United Kingdom

Abstract

Background: Segmentation performance with complete imaging may conceal sensitivity to sequence loss or contrast changes. **Objective:** We demonstrate risk-screened multimodal attribution (RSMA), a finite-library audit combining region-specific coalition utilities from independently screened predictors. **Methods:** A newly implemented simulation generated 320 two-dimensional MRI-like phantoms with four channels and three disjoint lesion regions. Four pixel-level classifiers were fitted on 160 phantoms, screened on 60, and evaluated on 100 held-out phantoms. Exact Shapley values were computed from 16 input subsets; 1,500 paired case bootstrap draws quantified attribution and rank uncertainty. **Results:** At the specified tolerance, all four candidates were retained and assigned unequal weights. RSMA had expected component-model macro Dice of 0.951 with complete inputs and 0.719 under mean single-channel deletion, versus 0.971 and 0.714 for the full-input baseline. Attribution drift under the specified shift was 0.047 (RSMA), 0.054 (baseline), and 0.046 (uniform pooling). Simultaneous intervals resolved 11 of 18 pairwise orderings. **Conclusion:** The experiment illustrates a trade-off between segmentation performance and attribution stability. RSMA incurred a full-input Dice cost relative to the validation-selected full-input model, while the mean single-deletion Dice was only modestly higher. The paired interval for shifted Dice relative to that baseline included zero. Uniform pooling produced lower drift, and its missing-input Dice was similar to RSMA. These findings do not support a universal advantage. At the primary tolerance no model was excluded; the comparison therefore primarily illustrates the effect of unequal risk-based weights. The narrower tolerance illustrates how the protocol can also select a strict subset. Model weights should be determined by deployment requirements rather than post hoc preference for an attribution profile.

Keywords: multimodal segmentation; Shapley value; missing modality; bootstrap; model multiplicity; synthetic phantom.

1 Introduction

Multisequence MRI provides complementary signals for delineating glioma-related abnormalities. Clinical acquisition guidance and segmentation benchmarks motivate examination of contrast-specific information rather than treating every channel as interchangeable [1,2]. Yet overlap scores alone cannot establish whether a predictor uses a stable combination of inputs. A model selected

*Corresponding author: zfguo278@ucl.ac.uk

for complete-input accuracy may depend disproportionately on a sequence that is degraded or unavailable at deployment.

Ren et al. [3] examined contrast-level Shapley explanations in a U-Net study using BraTS and a near-optimal model confidence-set construction. Their work motivates the present question, but its patient data and experimental outcomes are not reused here. Related work on model-class reliance and explanation uncertainty highlights the need to distinguish variability across acceptable models from uncertainty due to finite samples [4,5].

Beyond segmentation-specific research, recent developments in visual computing and medical imaging AI have explored alternative approaches to information representation and processing. Ding et al. [9] investigated decoder-free compressed-domain eye parsing through in-sensor analog convolution, illustrating an alternative approach to visual information processing under hardware constraints. In the medical imaging domain, Xi et al. [10] introduced a radiographic world model that learns shared latent representations for clinical reasoning and evidence generation. Although these studies address different imaging tasks, they provide a broader context for investigating how input representations relate to downstream predictive behavior. In multisequence MRI segmentation, a specific question is whether regional feature attribution remains stable when input availability and contrast characteristics change.

We introduce a finite-library audit, termed risk-screened multimodal attribution (RSMA), that addresses a different estimand from the confidence-set approach of Ren et al. [3]: screen models using a prespecified mixture of complete and missing-input risk, retain region-level rather than pooled explanations, and report simultaneous pairwise ordering intervals. The objective is not to recover an unknown population-optimal network but to describe a specified, validation-selected mixture of predictors and its behavior under controlled input stress. A synthetic phantom experiment enables an end-to-end demonstration without claiming clinical validation. The contribution is the integration of validation-based risk screening, region-specific coalition attribution and simultaneous pairwise inference in a single audit protocol; it is not a claim of a new Shapley estimator or a guarantee of population-optimal explanations.

2 Materials and methods

2.1 Synthetic phantom generation and split

A fixed random seed (20261126) generated 320 independent 32×32 pixel phantoms. A circular brain mask surrounded randomly translated and rotated elliptical lesions with perturbed boundaries. Three disjoint regions represented enhancing tumor (ET), necrotic core (NC), and edema (ED); these names denote simulation labels, not histopathological findings. Each pixel received one of four class-dependent intensity vectors, channel-specific gain and offset, correlated low-frequency noise, independent Gaussian noise, and mild spatial smoothing. Figure 1 summarizes the study partitions. The reference masks are generated mathematically and therefore do not represent human annotations.

The four channels are denoted T1ce, T1, FLAIR, and T2. Intensities are arbitrary normalized units rather than MRI acquisition physics. In this order, the class-mean vectors are $(0, 0, 0, 0)$ for normal tissue, $(1.6, 0.15, 0.55, 0.2)$ for ET, $(-0.1, -1.25, 0.2, 1.25)$ for NC, and $(0.1, 0.05, 1.5, 0.65)$ for ED. The circular brain radius is 14 pixels, and the lesion center is translated independently along each axis by a uniform draw from $[-3, 3]$ pixels. Ellipse semi-axes are uniform on $[6, 8]$ and $[5, 7]$ pixels, with rotation uniform on $[0, 2\pi)$. In rotated, axis-normalized coordinates, the radial distance is divided by $1 + 0.07\sin(3\theta + \alpha) + 0.04\cos(5\theta + \beta)$, with independent uniform phases on $[0, 2\pi)$. Radial thresholds 0.38, 0.72 and 1 define the nested NC, ET and ED labels (assigned disjointly). For each phantom and channel, gains and offsets are independent normal

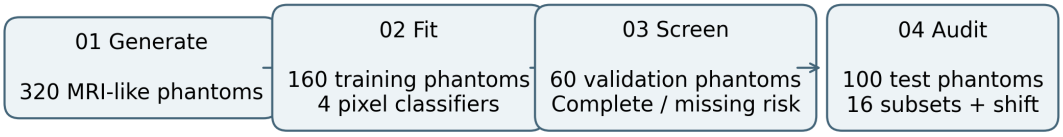

Figure 1: Workflow of the synthetic phantom study. Model fitting, validation screening and final testing use disjoint phantom IDs.

draws with means 1 and 0 and SDs 0.10 and 0.12. Independent standard-normal spatial fields are Gaussian-smoothed (SD 2 pixels), centered and scaled to unit spatial SD; a shared field is multiplied by 0.18 and channel-specific fields by 0.12. Independent Gaussian noise (SD 0.30) is then added and each channel is smoothed (SD 0.45 pixels; reflect boundaries). Pixels outside the brain mask are set to zero. Contrast preferences are partly known by construction; agreement cannot establish clinical reasoning.

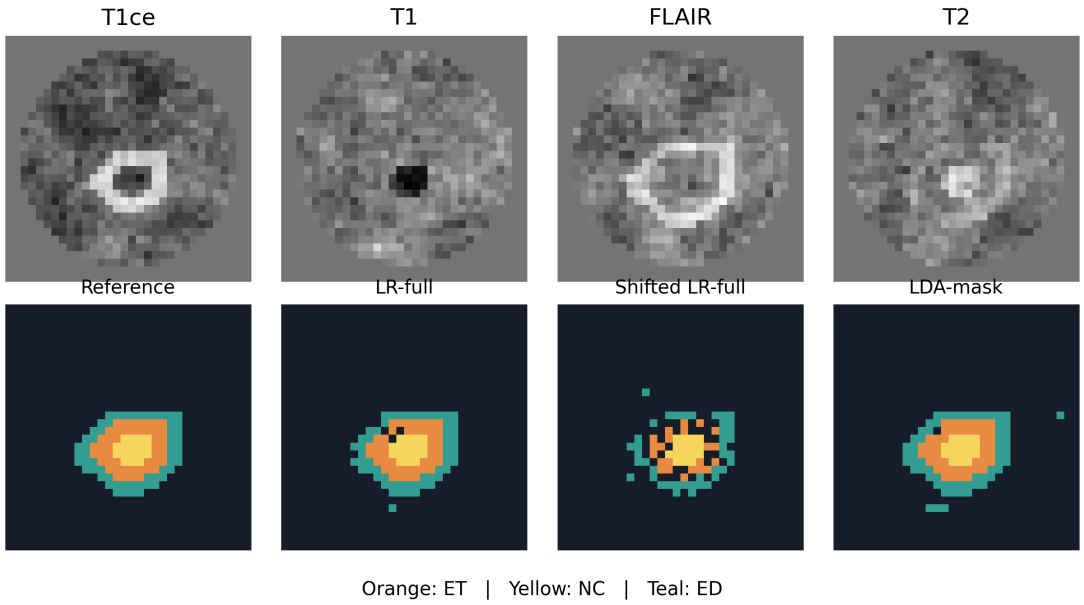

Figure 2: Example synthetic MRI-like phantom: four generated channels (top) and reference and predicted categorical masks (bottom). The shifted-input panel shows the baseline prediction after attenuation and noise addition; LDA-mask is the next highest-weight candidate after LR-full. These are not patient images.

The fixed partitions contain 160 training, 60 validation, and 100 test phantoms in generation order (IDs 0–159, 160–219 and 220–319). No test phantom is used for model fitting or validation-based weighting. The original generator was unavailable; the parameters specified here define an independent implementation of the documented protocol, not a reproduction of earlier numerical results.

2.2 Candidate predictors and missing-input interventions

Four lightweight classifiers were used so the entire experiment can run on a CPU: multinomial logistic regression trained on complete features (LR-full), logistic regression trained with channel

masking (LR-mask), linear discriminant analysis trained with masking (LDA-mask), and Gaussian naïve Bayes trained with masking (GNB-mask). These are pixel classifiers, not U-Net or transformer models. Each training phantom contributed 400 sampled in-mask pixels (240 normal, 65 ET, 35 NC, 60 ED; replacement allowed). In the masked training set, half the sampled pixels use all four channels and the remainder use a uniformly drawn nonempty subset. Unavailable intensities are set to zero. The same intervention is applied to all candidate models during testing; no explicit availability indicator is used. The training sampler uses seed 20261127 and samples pixels with replacement within each class. Logistic regression uses L2 regularization ($C = 1$), the L-BFGS solver, a maximum of 1,000 iterations and tolerance 10^{-7}; LDA uses the SVD solver and Gaussian naive Bayes uses variance smoothing 10^{-9}. No spatial coordinates are supplied to the classifiers.

For phantom i, region r, candidate j, and available channel set S, let $D_{ijr}(S)$ denote the Dice coefficient. Labels outside the circular mask are fixed to background. The denominator is floored at one to avoid division by zero; when both masks are empty, the resulting Dice value is zero under this convention.

$$D_{ijr}(S) = \frac{2\,|P_{ijr}(S) \cap G_{ir}|}{\max(1, |P_{ijr}(S)| + |G_{ir}|)} \tag{1}$$

2.3 Risk-screened weighting

Let F_j and H_j be validation mean Dice over cases and regions with complete input and averaged over the four single-channel deletion conditions, respectively. The screening risk balances these equally. The tolerance $\delta = 0.04$ and temperature $T = 0.025$ are illustrative design constants, not clinically calibrated thresholds.

$$R_j = 1 - (0.5F_j + 0.5H_j), \quad E = \{j : R_j \le \min_k R_k + \delta\} \tag{2}$$

$$w_j = \frac{\mathbf{1}(j \in E)\exp[-(R_j - \min_k R_k)/T]}{Z} \tag{3}$$

Here, Z is the sum of unnormalized eligible weights. We compare this fixed weighting with a single candidate maximizing full-input validation Dice and with uniform weights over all four candidates. Weighted Dice denotes expected performance of a randomly selected component, not Dice of an ensembled mask. The same weights are used for attributions; test performance is never used to revise them.

2.4 Exact regional coalition attribution

Using four channels, all 16 subsets are evaluated, including the empty subset (all four intensities set to zero). The contribution of channel c is the weighted marginal gain over subsets not containing c [6]. The following formula retains the region and case indices.

$$\phi_{ijrc} = \sum_{S \subseteq M \setminus \{c\}} \frac{|S|!(3 - |S|)!}{4!} \left[D_{ijr}(S \cup \{c\}) - D_{ijr}(S)\right] \tag{4}$$

The sum is over $S \subseteq M$ excluding c, with M the four-channel set. Exact computation satisfies $\sum_c \phi_{ijrc} = D_{ijr}(M) - D_{ijr}(\varnothing)$. This identity provides a numerical check for both clean and shifted cases. The weighted case contribution is $a_{irc} = \sum_j w_j \phi_{ijrc}$, and the target is its mean over the held-out case distribution. Linearity gives the same result as applying Shapley attribution to the

weighted coalition utility. Negative contributions are allowed; attribution is intervention-specific, not causal.

2.5 Bootstrap inference and stress testing

We draw 1,500 paired bootstrap samples of 100 test IDs (seed 777), retaining all regions, modalities and model outputs together. For each draw b, calculate the mean attribution vector $\bar{a}^{*(b)}$. A common 95th percentile radius of the largest absolute centered deviation gives a simultaneous approximate band for the 12 mean contributions.

$$\begin{aligned} q &= Q_{0.95}\left(\max_{r,c} |\bar{a}^{*}_{rc} - \bar{a}_{rc}|\right), \\ \mathrm{CI}_{rc} &= [\bar{a}_{rc} - q, \bar{a}_{rc} + q]. \end{aligned} \tag{5}$$

An analogous maximum over the 18 regional pairwise differences yields a separate simultaneous band. An ordering is resolved only if its interval excludes zero. These bootstrap intervals are conditional on the fitted candidate library and weights, and are not finite-sample guarantees. The shift multiplies T1ce by 0.65 and FLAIR by 0.70 and adds Gaussian noise with SD 0.30 and 0.25, respectively, preserving the test masks. Independent perturbation noise is generated with seed 20261128 before coalition masking. Drift is the mean absolute clean-to-shift attribution difference across cases, regions and channels. We also report marginal paired percentile intervals for mean differences in full, missing, shifted Dice and drift between RSMA and each comparator, using the same bootstrap draws; these exploratory contrasts are not adjusted for multiplicity.

3 Results

3.1 Candidate performance and screening

Table 1 summarizes candidate performance and validation-derived weights. The validation-selected full-input model was LR-full. Its test macro Dice was 0.971; LDA-mask had the highest held-out mean single-deletion Dice (0.726). At $\delta = 0.04$, all four candidates were eligible; their weights in table order were 0.365, 0.155, 0.358, 0.122. Thus the primary RSMA analysis is a risk-weighted pool rather than a strict subset of the library. Eligibility does not imply statistical equivalence.

Table 1: Held-out candidate performance and validation-derived weights.

Candidate	ET	NC	ED	Full	Missing[1]	Shift[2]	Risk[3]	Weight
LR-full	0.971	0.987	0.957	0.971	0.714	0.835	0.159	0.365
LR-mask	0.935	0.945	0.913	0.931	0.709	0.830	0.180	0.155
LDA-mask	0.958	0.974	0.940	0.957	0.726	0.842	0.159	0.358
GNB-mask	0.868	0.940	0.872	0.893	0.724	0.807	0.186	0.122

All values are from synthetic phantoms. [1] Mean across four single-channel deletion conditions. [2] Complete-input Dice under the specified contrast shift. [3] Validation risk (Eq. 2). Dice is macro-averaged over regions and cases.

3.2 Region-level attribution

The risk-screened mean attribution surface (Figure 3) showed distinct region-specific patterns. T1ce dominated ET (0.948 Dice units), and FLAIR dominated ED (0.873). In NC, T1 and T2 contributed 0.536 and 0.404, respectively. These patterns are consistent with the programmed

signal profiles. Negative contributions can occur because the coalition utility is nonlinear and a channel can change predictions in an unhelpful way. The maximum absolute Shapley efficiency residual across validation, clean test and shifted test cases was 3.33×10^{-16}.

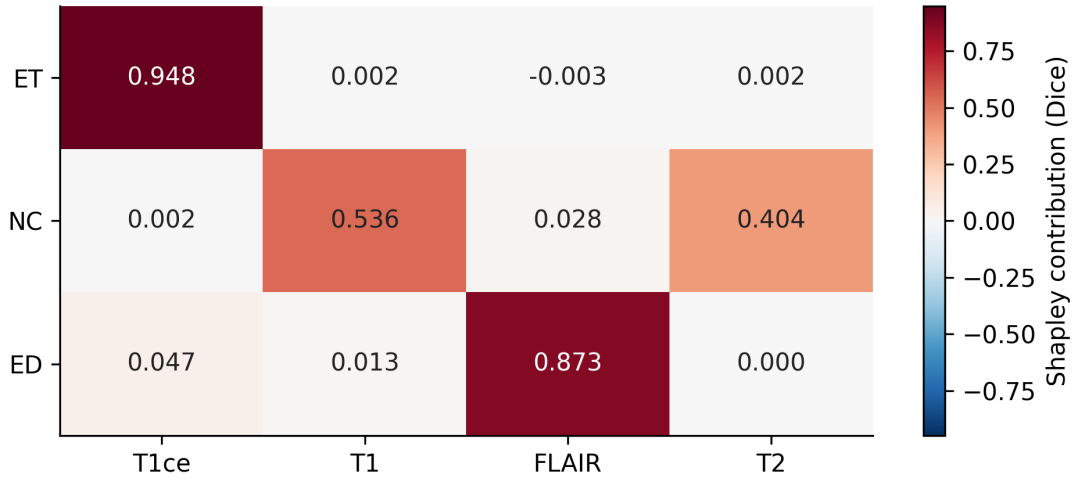

Figure 3: Mean exact Shapley contributions by lesion label and channel after risk screening, calculated on synthetic test phantoms. The colour scale denotes signed Dice contributions, not image intensity.

The across-case mean alone does not reveal whether a contrast ordering is stable. We therefore retain the full case-level attribution tensor and report simultaneous uncertainty for its components and pairwise differences rather than pooling across regions.

3.3 Simultaneous uncertainty and partial ordering

Table 2 reports region-specific means and simultaneous intervals. The bootstrap maximum-deviation radius was 0.0369 Dice units across the 12 mean contributions; the separate pairwise radius was 0.0674 across 18 differences. Eleven differences excluded zero. T1ce exceeded all alternatives for ET and FLAIR exceeded all alternatives for ED. For NC, both T1 and T2 exceeded T1ce and FLAIR, and T1 exceeded T2 (difference 0.132; simultaneous interval [0.065, 0.199]). All other pairwise comparisons were unresolved, not proven equal.

Table 2: Region-specific mean attribution and simultaneous 95% bootstrap bands.

Region	Channel	Clean mean	Lower	Upper	Shift mean
ET	T1ce	0.948	0.911	0.985	0.712
ET	T1	0.002	-0.035	0.038	-0.004
ET	FLAIR	-0.003	-0.040	0.034	0.023
ET	T2	0.002	-0.035	0.039	0.005
NC	T1ce	0.002	-0.034	0.039	-0.000
NC	T1	0.536	0.499	0.573	0.524
NC	FLAIR	0.028	-0.009	0.065	0.021
NC	T2	0.404	0.367	0.441	0.404
ED	T1ce	0.047	0.010	0.084	0.013
ED	T1	0.013	-0.024	0.050	0.007
ED	FLAIR	0.873	0.836	0.910	0.739
ED	T2	0.000	-0.037	0.037	0.056

All values are from synthetic phantoms. Intervals are approximate simultaneous 95% bootstrap bands for the clean mean attributions; the shifted means are descriptive.

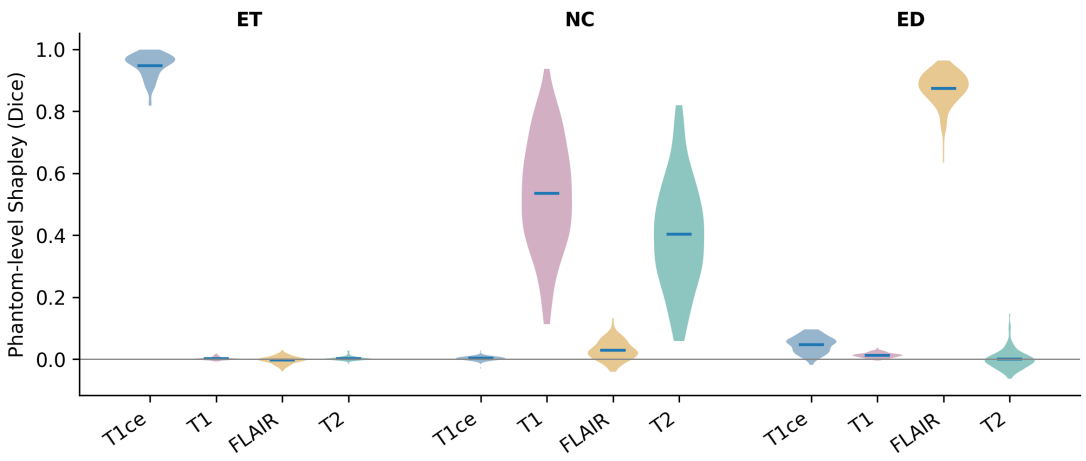

Figure 4: Distribution of phantom-level Shapley contributions across 100 independent synthetic test phantoms. Groups correspond to ET, NC and ED; short horizontal marks indicate means.

Figure 4 illustrates how the estimated mean is only one summary of a heterogeneous case distribution. The plotted spread describes individual phantoms and is not itself a confidence interval for the mean.

3.4 Risk–performance trade-off and missing-input robustness

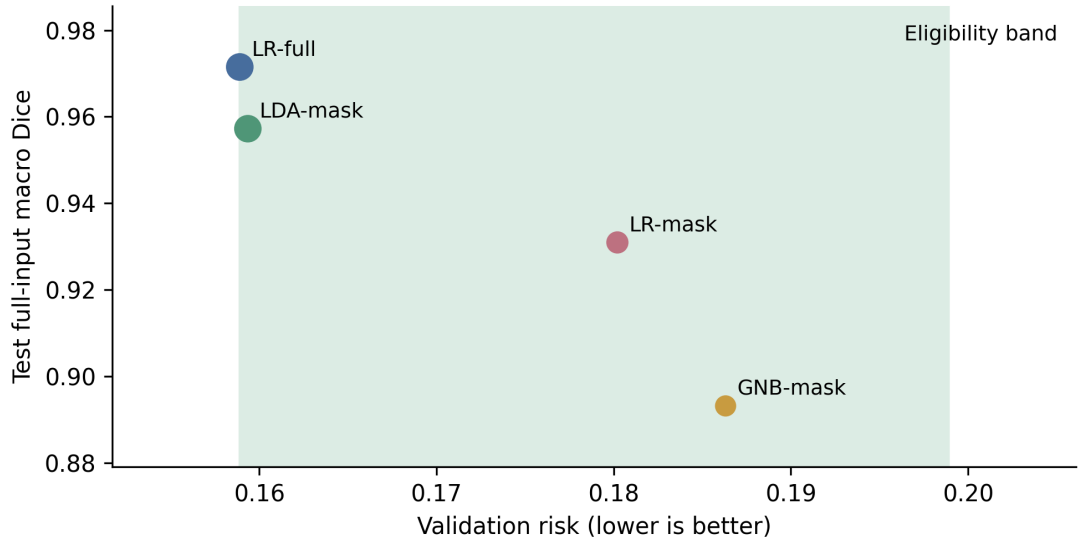

Figure 5: Validation risk versus complete-input test performance in the synthetic experiment. Symbol size is proportional to the selected weight plus a constant visibility offset; the shaded region denotes the eligibility band.

The four candidates span different performance profiles (Figure 5). RSMA reduced full-input macro Dice from 0.971 to 0.951 and changed the mean single-deletion score from 0.714 to 0.719. The small absolute missing-input gain and the full-input performance cost should be considered together.

Figure 6 shows why reporting only an average deletion score is insufficient: removal of a different channel can affect different regions and candidates. The synthetic setting measures response

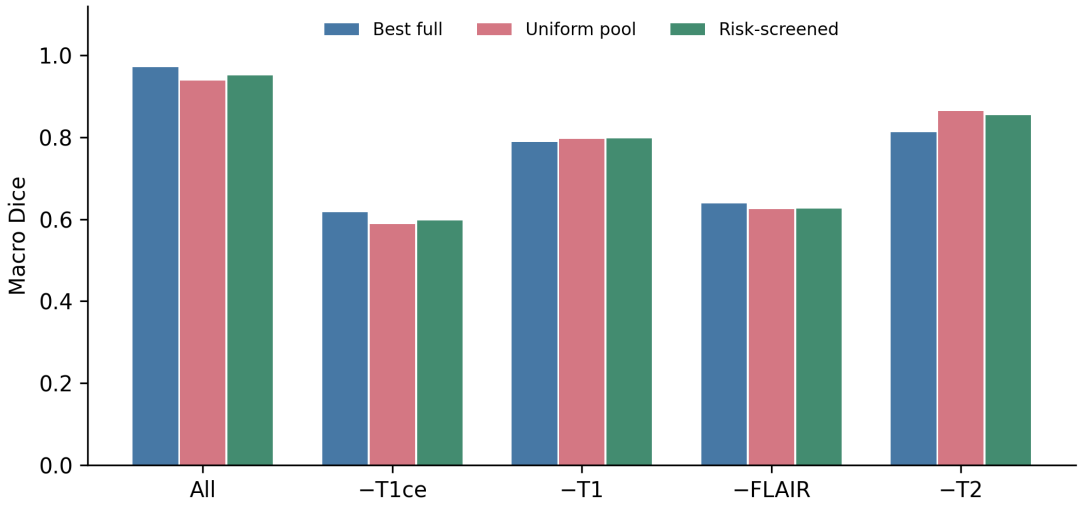

Figure 6: Expected component-model macro Dice for complete input and each single-channel deletion on synthetic test phantoms. Weights are fixed across conditions; no models are retrained for the deletions.

to a fixed zero-mask intervention. In practice, image imputation, missingness indicators or modality-specific training could produce different behavior.

3.5 Attribution drift and sensitivity analysis

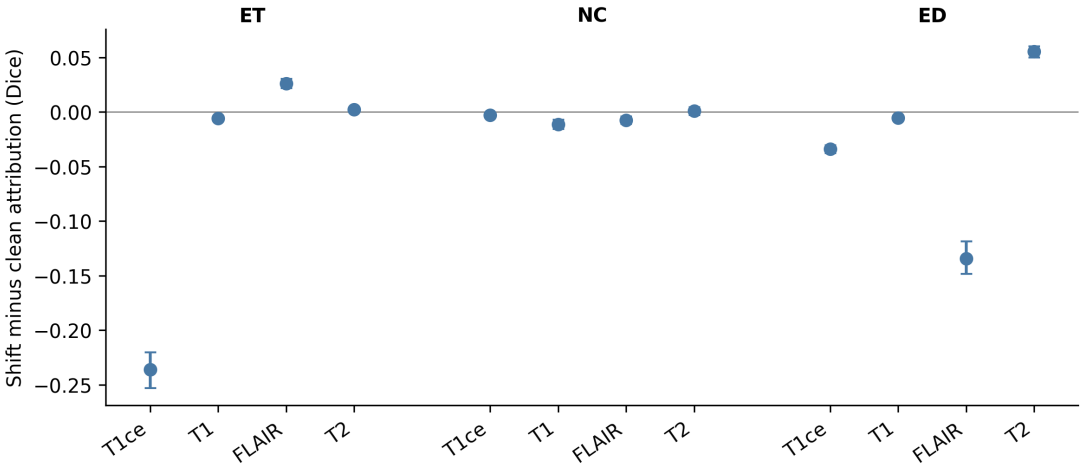

Figure 7: Paired changes in RSMA attribution after attenuation and noise on synthetic test phantoms. Error bars are marginal percentile bootstrap intervals for mean paired changes and are not multiplicity-adjusted.

Table 3 summarizes the three weighting strategies. The perturbation reduced average T1ce attribution to ET from 0.948 to 0.712 and FLAIR attribution to ED from 0.873 to 0.739. RSMA drift was 0.047, compared with 0.054 for the baseline and 0.046 for uniform pooling. Relative to the full-input baseline, the RSMA mean differences (marginal 95% paired percentile intervals) were −0.0210 [−0.0241, −0.0179] for full Dice, 0.0045 [0.0012, 0.0078] for missing Dice, −0.0017 [−0.0057, 0.0024] for shifted Dice, and −0.0076 [−0.0083, −0.0070] for drift. Compared with uniform pooling, the corresponding differences were 0.0123 [0.0110, 0.0137], 0.0004 [−0.0008, 0.0017], 0.0048 [0.0033, 0.0063], and 0.0010 [0.0008, 0.0012]. These exploratory intervals do not establish simultaneous between-strategy superiority.

Table 3: Comparison of attribution weighting strategies.

Strategy	Full	Missing	Shift	Drop [95% CI][1]	Drift
Best full	0.971	0.714	0.835	0.257 [0.252, 0.263]	0.054
Uniform pool	0.938	0.718	0.828	0.220 [0.214, 0.226]	0.046
Risk-screened	0.951	0.719	0.833	0.232 [0.226, 0.237]	0.047

All values are from synthetic phantoms. [1] Paired case-bootstrap percentile interval for complete-input minus mean single-deletion Dice. The intervals quantify within-strategy drops, not between-strategy differences.

Table 4: Sensitivity to the validation eligibility tolerance δ.

δ	Retained	Effective N[1]	Full	Missing	Drift
0.00	1	1.000	0.971	0.714	0.054
0.02	2	2.000	0.964	0.720	0.048
0.04	4	3.331	0.951	0.719	0.047
0.08	4	3.331	0.951	0.719	0.047
0.15	4	3.331	0.951	0.719	0.047

All values are from synthetic phantoms. [1] Effective number of components: $1/\sum_j w_j^2$. The temperature and held-out test set remain fixed; this is a descriptive sensitivity analysis, not a tuning procedure. At $\delta = 0.02$, two candidates are retained; at $\delta \geq 0.04$, all four are retained for the tested grid.

Table 4 reports sensitivity to the eligibility tolerance. Two candidates are retained at 0.02, whereas all four are retained at 0.04, 0.08 and 0.15. The primary threshold is unchanged; the sensitivity results should not be used to choose a threshold retrospectively on the test set.

4 Discussion

This proof of concept addresses a different estimand from the source study [3]. Rather than asserting that a confidence set contains a population-optimal model, RSMA summarizes the coalition utility of a validation-selected finite predictor mixture. Its methodological value is the explicit separation of predictive screening, case-level uncertainty, and stress testing. The approach can be used when the candidate library and channel intervention are specified; it does not require global optimization over a neural-network parameter space.

The experiment illustrates a trade-off between segmentation performance and attribution stability. Although RSMA improved the missing-input average relative to the complete-input baseline, its complete-input Dice declined and its shifted Dice remained below the baseline. Uniform pooling produced lower attribution drift, while the risk-screened strategy gave better absolute full and missing Dice. Consequently, the results suggest a trade-off rather than a universal advantage. The 0.010 difference in mean single-deletion Dice is descriptive; the reported intervals for performance drops do not establish a statistically significant between-strategy difference. Model weights should be determined by deployment requirements, not by post hoc preference for an attractive attribution profile.

Simultaneous pairwise intervals provide an alternative to forcing a total order of MRI channels. Some low-contribution channel comparisons remain unresolved, even with the simplified generator. The intervals account only for held-out case sampling with fixed fits and weights. They do not cover variation from refitting, validation selection, scanner differences or unknown population shifts. Repeated training and external site validation are necessary before stronger reliability claims are justified. Only one generator parameterization, one partition and one specified contrast shift are evaluated; sensitivity to the risk-mixing coefficient, temperature and alternative missing-input interventions remains untested.

4.1 Limitations and planned clinical study

The simulation uses simplified two-dimensional lesion shapes, arbitrary intensity distributions and pixel classifiers. Neither real MRI physics, clinical annotation uncertainty, class prevalence nor the spatial complexity of gliomas is represented faithfully. The apparent correspondence of contrast and label is partly encoded by the generator. No human participants, real patient images, clinical annotations, nnU-Net or Swin UNETR training, or actual BraTS experiments are included. A real study should use predeclared patient-level train/validation/test splits, independent sites, several seeds per architecture, and the same coalition intervention across models. Candidate libraries could include nnU-Net [7] and Swin UNETR [8], with modality dropout and conventional training represented separately. Compare full and missing-input Dice, HD95 and calibration where appropriate; collect neuroradiologist rankings independently if clinical agreement is a primary endpoint. Estimate joint uncertainty by repeated fitting and hierarchical or site-aware resampling, and lock thresholds before external testing. No clinical performance value is supplied for that future study.

5 Conclusion

Risk-screened multimodal attribution specifies an audit of region-specific feature use under candidate-model multiplicity and missing inputs. The independently implemented synthetic study demonstrates how risk weighting, exact coalition utilities and simultaneous inference can be combined, and illustrates accuracy–stability trade-offs without establishing superiority. The findings are conditional on the simulated generator and classifiers. Clinical validation remains a necessary subsequent step.

Declarations

Ethics approval and consent to participate. Not applicable. This study uses mathematically generated synthetic phantoms and does not involve human participants, human tissue, animal subjects or identifiable patient data.

Consent for publication. Not applicable to the synthetic images and results reported here.

Data and code availability. The accompanying reproducibility archive contains the newly implemented Python script, configuration, dependency versions, generated phantom arrays, coalition utilities, all 18 pairwise contrasts, strategy contrasts and figure-generation routines. The files enable regeneration of the experiment described here. No public repository or persistent identifier has yet been assigned.

Competing interests. The authors declare no competing interests.

Funding. This study received no external funding.

Authors' contributions. Zifan Guo: methodology. Piao Ti: experiments.

References

1. Ellingson BM, Bendszus M, Boxerman J, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. Neuro-Oncology. 2015;17(9):1188–1198. doi:10.1093/neuonc/nov095.

2. Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on

brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021).

3. Ren T, Low D, Xiang R, Jaengprajak P, Rivera JH, Olson R, Ruzevick J, Kurt M. Can you trust your model? Constructing uncertainty approximations guaranteeing validity of glioma segmentation explanations. Proceedings of Machine Learning Research. 2026;315:1407–1421. https://proceedings.mlr.press/v315/ren26a.html.

4. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20(177):1–81. https://www.jmlr.org/papers/v20/18-760.html.

5. Marx C, Park Y, Hasson H, Wang Y, Ermon S, Huan L. But are you sure? An uncertainty-aware perspective on explainable AI. Proc AISTATS. 2023;206:7375–7391. https://proceedings.mlr.press/v206/marx23a.html.

6. Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, eds. Contributions to the Theory of Games II. Princeton University Press; 1953:307–317.

7. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–211. doi:10.1038/s41592-020-01008-z.

8. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer; 2022. arXiv:2201.01266.

9. Ding H, Tabrizchi S, Roohi A. A2CD-Gaze: Decoder-Free Compressed-Domain Eye Parsing via In-Sensor Pre-ADC Analog Convolution. In: ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED). 2026:1–7.

10. Xi S, Hu S, Wang S, Safari M, del Balzo L, Karim EU, Hu M, Zhang K, Wang T, Weichselbaum RR, et al. A radiographic world model for clinical reasoning and evidence generation. arXiv:2609.07719 (2026). https://arxiv.org/abs/2609.07719.

Appendix A. Reproducibility checklist

The archive contains `simulate.py`, `requirements.txt`, `README.md`, and the `results/` directory. Run `python simulate.py` in the recorded environment to regenerate the 320 phantoms, coalition utilities, CSV summaries, and seven figures. The generator seed is 20261126, and the bootstrap seed is 777. The data arrays preserve phantom IDs; coalition indices use bit order T1ce, T1, FLAIR, T2 (least to most significant). Table values are rounded to three decimals and the two bootstrap radii to four.

The reported results concern only the specified synthetic generator and finite classifier library. The methodological contribution is a proof of concept; novelty relative to the full literature and generalisation to clinical MRI remain to be established.