RELIABLE NEUROIMAGING MODEL INTERPRETATION VIA CLOSED-FORM ROI INFLUENCE POSTERIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neuroimaging-based machine learning models are increasingly used to study brain disorders and aging, yet their interpretations are often reported as point importance maps over brain regions. In region-of-interest (ROI) analyses of structural MRI and fMRI, limited sample sizes, heterogeneous cohorts, and training uncertainty can change which regions appear influential. Standard attribution methods rarely quantify whether a regional effect is stable. We propose **Closed-Form Influence Posterior** (CFIP), a probabilistic framework for reliable neuroimaging model interpretation. CFIP defines regional influence as the logit change induced by intervening on an ROI, models parameter uncertainty with a Laplace approximation, and propagates it through a linearized influence map. This construction gives a closed-form posterior over regional effects. The pushforward is exact for an affine last layer with fixed features under the Gaussian parameter approximation, and extends to all layers through a tangent model. The influence posterior supports reliability assessment and decomposes attribution instability into model uncertainty and data heterogeneity. We validate CFIP across structural MRI and fMRI benchmarks, including Alzheimer's disease and aging settings. Comparisons with Bayesian baselines, HMC, and synthetic benchmarks evaluate influence rankings and their uncertainty. CFIP produces faithful regional explanations, uncertainty estimates that track observed attribution variability, and sampling-free influence inference, providing a practical reliability layer for neuroimaging model interpretation. Codes are available at `https://anonymous.4open.science/r/CFIP-7BE1/`.

1 INTRODUCTION

Neuroimaging-based machine learning models are increasingly used to study brain disorders, cognitive impairment, and aging, but their utility depends on whether model decisions can be interpreted in meaningful brain regions (Jack et al., 2008; Petersen et al., 2010; Cuingnet et al., 2011; Rathore et al., 2017). Region-of-interest (ROI) representations provide a natural interface for this goal: structural MRI ROIs summarize anatomical measurements such as cortical thickness and volume, while fMRI connectivity represents functional relationships among brain regions (Fischl, 2012; Desikan et al., 2006; Cho et al., 2012; Biswal et al., 1995; van den Heuvel & Hulshoff Pol, 2010; Power et al., 2011; Smith et al., 2013). Across these settings, region-level explanations align better with neuroimaging analysis than generic pixel- or voxel-level saliency.

Most post-hoc attribution methods, such as gradient saliency, Integrated Gradients, LIME, and SHAP, produce point importance estimates for one trained model (Simonyan et al., 2014; Sundararajan et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017). Such point explanations rarely indicate whether a reported regional effect is stable under training randomness, limited data, or cohort heterogeneity. This is problematic in neuroimaging, where datasets are often small, diagnostic labels can be noisy, and saliency maps can react inappropriately to input transformations or remain insensitive to model randomization (Adebayo et al., 2018; Kindermans et al., 2019; Ghorbani et al., 2019). MC-dropout and ensembles model predictive uncertainty through stochastic passes or independently fitted models (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). Recomputing attributions across these models also estimates explanation variability, but requires repeated explanation computation.

We propose **Closed-Form Influence Posterior** (CFIP), a probabilistic framework for reliable regional neuroimaging model interpretation. CFIP defines the influence of an ROI as the model-output change

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.

induced by intervening on that region, approximates model uncertainty with a Laplace posterior, and propagates this uncertainty through a linearized ROI influence functional (MacKay, 1992; Ritter et al., 2018; Daxberger et al., 2021). The resulting closed-form posterior over ROI influences turns a single importance map into a distributional view of regional effects, supporting uncertainty-aware interpretation, reliability assessment, and decomposition of attribution instability into model uncertainty and data heterogeneity.

Our main experiments use last-layer Laplace with a fixed feature extractor. The logit is affine in the last-layer parameters, so the Gaussian parameter posterior gives an exact Gaussian influence distribution. We also introduce a full-network tangent variant that incorporates all-layer gradient features while preserving closed-form inference. We analyze its approximation error and evaluate its accuracy against nonlinear-network samples.

We validate CFIP across multi-cohort structural MRI and fMRI benchmarks, including Alzheimer's disease cohorts such as ADNI and OASIS and aging-related benchmarks such as HCP-Aging (Jack et al., 2008; Petersen et al., 2010; LaMontagne et al., 2019; Bookheimer et al., 2019; Harms et al., 2018). Our contributions are: **(1)** a closed-form influence posterior for anatomical and functional neuroimaging ROIs under a Laplace-linearized construction; **(2)** a full-network tangent extension with an approximation analysis; **(3)** reliability assessment through influence uncertainty and its decomposition into model uncertainty and data heterogeneity; and **(4)** multi-cohort structural MRI and fMRI validation, with Bayesian baseline, HMC, and synthetic comparisons that assess faithfulness, uncertainty, and computational efficiency.

2 Related Work

Regional neuroimaging and point attributions. ROI representations summarize structural MRI and fMRI into anatomically or functionally meaningful units (Fischl, 2012; Rathore et al., 2017; Biswal et al., 1995; Power et al., 2011). Saliency, Integrated Gradients, LIME, and SHAP assign importance to these features for a trained predictor (Simonyan et al., 2014; Sundararajan et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017). Their sensitivity to inputs and model parameters motivates assessing explanation reliability (Adebayo et al., 2018; Kindermans et al., 2019; Ghorbani et al., 2019).

Uncertainty-aware explanations. MC-dropout and ensembles estimate variability through stochastic passes or retrained models (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). BayesGrad averages gradients over dropout samples; distributional SHAP studies explanations under uncertain population distributions (Akita et al., 2018; Cifuentes et al., 2024). CFIP combines a Laplace parameter posterior (MacKay, 1992; Ritter et al., 2018; Daxberger et al., 2021) with ROI intervention to obtain an analytic influence distribution and separate model uncertainty from data heterogeneity.

3 Background and Problem Setup

We study binary neuroimaging prediction from regional brain features. Let $\mathcal{D}$ be the training set and $x = (x_1, \ldots, x_m) \in \mathbb{R}^{m \times d}$ denote one subject's region-of-interest (ROI) representation, where x_k is the feature vector for ROI k. In structural MRI, x_k contains anatomical measurements such as cortical thickness or volume; in fMRI, it can encode the connectivity profile of ROI k (Fischl, 2012; Cho et al., 2012; Biswal et al., 1995; van den Heuvel & Hulshoff Pol, 2010). A model f_θ maps x to a scalar logit, which is converted to a class probability by a sigmoid.

ROI-level interpretability asks how each brain region contributes to a model decision. Standard attribution methods return a single importance score per ROI (Simonyan et al., 2014; Sundararajan et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017), but such point estimates do not indicate whether a regional effect is reliable under model uncertainty or data heterogeneity.

We adopt a Bayesian view in which the model parameters, or a selected stochastic parameter block such as the last layer, have posterior $p(\theta \mid \mathcal{D})$. Since exact Bayesian inference is intractable for neural models, we use a Laplace approximation around a trained estimate $\hat{\theta}$, yielding a Gaussian posterior (MacKay, 1992; Ritter et al., 2018; Daxberger et al., 2021). CFIP propagates this posterior to obtain a distribution over ROI influences.

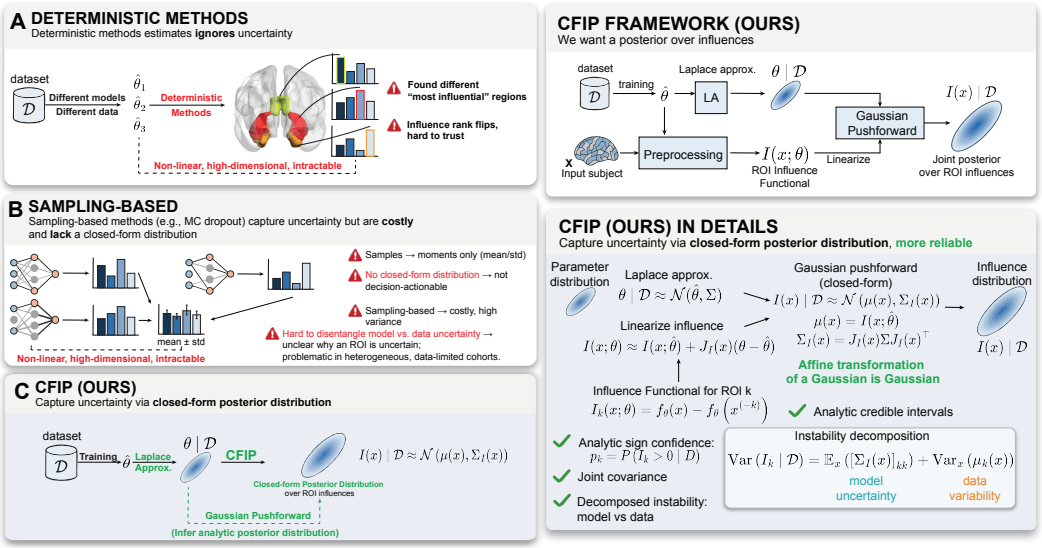


Figure 1: Overview of CFIP. ROI intervention, Laplace approximation, and linearization give a closed-form Gaussian influence posterior for regional reliability assessment and instability decomposition.

4 METHOD

CFIP converts a deterministic ROI importance map into a posterior distribution over regional influences. We define an intervention-based ROI effect, propagate parameter uncertainty to its closed-form posterior, and decompose attribution instability into model uncertainty and data heterogeneity. We then extend the construction to all layers of the network.

4.1 ROI INFLUENCE FUNCTIONAL FOR REGIONAL EFFECT MEASUREMENT

Let $x^{(-k)} = \mathcal{M}_k(x)$ denote a fixed intervention that replaces ROI k with a baseline. For structural MRI, $x^{(-k)} = (x_1, \ldots, x_{k-1}, b_k, x_{k+1}, \ldots, x_m)$, where b_k is a standardized zero or training-set mean. For symmetric fMRI connectivity, $\mathcal{M}_k$ replaces edges incident to region k in both its row and column. We define ROI influence as the intervention-induced logit change:

$$I_k(x;\theta) = f_\theta(x) - f_\theta(x^{(-k)}), \qquad I(x;\theta) = [I_1(x;\theta), \ldots, I_m(x;\theta)]^\top. \tag{1}$$

Positive values indicate ROIs that support the positive-class prediction; negative values indicate opposing ROIs. The effect measures model reliance under the chosen baseline.

4.2 TRACTABLE INFERENCE FROM PARAMETER TO INFLUENCE DISTRIBUTION

Given the parameter posterior $p(\theta \mid \mathcal{D})$, the induced influence distribution is the pushforward

$$p(I(x) \mid \mathcal{D}) = \int \delta\big(I(x) - I(x;\theta)\big) p(\theta \mid \mathcal{D})\, d\theta. \tag{2}$$

CFIP makes this distribution tractable by combining a Gaussian parameter approximation with a first-order influence map.

Laplace parameter posterior. With prior $\theta \sim \mathcal{N}(0, \lambda^{-1} I)$ and $\lambda > 0$, training gives

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)) + \frac{\lambda}{2}\|\theta\|_2^2. \tag{3}$$

The Laplace approximation is

$$p(\theta \mid \mathcal{D}) \approx q(\theta) = \mathcal{N}(\hat{\theta}, \Sigma), \qquad \Sigma \approx H^{-1}, \tag{4}$$

where H is a positive definite curvature approximation to the negative log-posterior (MacKay, 1992; Ritter et al., 2018; Daxberger et al., 2021). Our main experiments use last-layer diagonal GGN Laplace. For $f_\theta(x) = w^\top h_\psi(x) + b$, the extractor ψ is fixed and $\theta = (w, b)$ is stochastic. With $\tilde{h}_i = [h_\psi(x_i)^\top, 1]^\top$ and $p_i = \sigma(f_{\hat{\theta}}(x_i))$, the precision is

$$\Lambda_{\text{diag}} = \lambda I + \sum_i p_i(1 - p_i)\,\text{diag}(\tilde{h}_i \odot \tilde{h}_i), \qquad \Sigma = \Lambda_{\text{diag}}^{-1}. \tag{5}$$

Linearized influence map. Around the trained model, define

$$\widetilde{I}(x;\theta) = I(x;\hat{\theta}) + J_I(x)(\theta - \hat{\theta}), \quad [J_I(x)]_{k,:} = \left(\nabla_\theta f_{\hat{\theta}}(x) - \nabla_\theta f_{\hat{\theta}}(x^{(-k)})\right)^\top. \tag{6}$$

This approximates $I(x;\theta)$ and is exact when the logit is affine in the stochastic parameters.

4.3 Closed-Form Posterior over ROI Influences

An affine transformation of a Gaussian is Gaussian. Under $q(\theta)$,

$$\widetilde{I}(x) \sim \mathcal{N}(\mu(x), \Sigma_I(x)), \qquad \mu(x) = I(x;\hat{\theta}), \quad \Sigma_I(x) = J_I(x)\Sigma J_I(x)^\top. \tag{7}$$

Each ROI therefore has marginal $\mathcal{N}(\mu_k(x), [\Sigma_I(x)]_{kk})$ and sign confidence

$$\mathbb{P}_q(\widetilde{I}_k(x) > 0) = \Phi\left(\frac{\mu_k(x)}{\sqrt{[\Sigma_I(x)]_{kk}}}\right), \tag{8}$$

for positive variance, where Φ is the standard Normal CDF. At zero variance it is $\mathbf{1}\{\mu_k(x) > 0\}$. The mean determines the ROI ranking, while the covariance describes uncertainty around these effects and supports joint analysis of regional influence maps.

Theorem 1 (Exact influence posterior for linear stochastic parameters)**.** *Let $f_\theta(x) = \langle \theta, \phi(x) \rangle$ for fixed features ϕ, and $A_{k,:}(x) = (\phi(x) - \phi(x^{(-k)}))^\top$. If $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma)$, then $I(x;\theta) = A(x)\theta \sim \mathcal{N}(A(x)\hat{\theta}, A(x)\Sigma A(x)^\top)$. Thus Eq.* (7) *is exact under the Gaussian parameter approximation for logistic regression and a last layer with fixed features.*

4.4 Decomposition of Attribution Instability

For a fixed training posterior and an input drawn from the empirical data distribution, the law of total variance gives

$$\text{Var}_q(\widetilde{I}_k \mid \mathcal{D}) = \underbrace{\mathbb{E}_x([\Sigma_I(x)]_{kk})}_{\text{model uncertainty}} + \underbrace{\text{Var}_x(\mu_k(x))}_{\text{data heterogeneity}}. \tag{9}$$

The first term measures parameter-induced attribution uncertainty; the second measures differences in subjects' mean regional effects. We obtain cohort rankings by averaging $|\mu_k(x)|$ over held-out subjects across folds; deletion and insertion rank each subject separately.

4.5 Full-Network Tangent CFIP

To include uncertainty from all layers, let θ contain every network parameter and define $\psi_{\hat{\theta}}(z) = \nabla_\theta f_{\hat{\theta}}(z)$. The tangent model $\tilde{f}_\alpha(z) = f_{\hat{\theta}}(z) + \psi_{\hat{\theta}}(z)^\top \alpha$, with $\alpha \sim \mathcal{N}(0, \Sigma_{\tan})$, gives the same closed-form Gaussian law in Eq. (7) using the all-parameter Jacobian. For layer blocks ℓ, r and influence sensitivities $a_{k,\ell} = \nabla_{\theta_\ell} f_{\hat{\theta}}(x) - \nabla_{\theta_\ell} f_{\hat{\theta}}(x^{(-k)})$,

$$\text{Var}(\widetilde{I}_k) = \sum_{\ell,r} a_{k,\ell}^\top [\Sigma_{\tan}]_{\ell r} a_{k,r}. \tag{10}$$

Every layer is included through its local sensitivity; a full covariance also retains cross-layer correlations. A diagonal covariance uses all coordinates while omitting these correlations. Appendix C derives the precision and posterior center from the original regularized objective.

Proposition 2 (Extractor contribution to influence uncertainty)**.** *Partition the all-parameter Gaussian precision and influence Jacobian into head and extractor blocks:* $H = \left(\begin{smallmatrix} C & B \\ B^\top & D \end{smallmatrix}\right) \succ 0$, $J = [J_h, J_e]$. *With* $S = D - B^\top C^{-1} B$ *and* $G = J_e - J_h C^{-1} B$,

$$\underbrace{JH^{-1}J^\top}_{\text{all parameters uncertain}} = \underbrace{J_h C^{-1} J_h^\top}_{\text{extractor fixed}} + \underbrace{GS^{-1}G^\top}_{\text{extractor contribution}} . \tag{11}$$

The additional term is positive semidefinite. The fixed-extractor comparison conditions the same Gaussian on zero extractor perturbation.

This decomposition shows how extractor uncertainty enters the influence posterior while preserving analytic inference. It compares parameter blocks within one Gaussian approximation; its proof is in Appendix A.

4.6 Approximation Analysis

The all-layer Gaussian law is exact for the tangent model. For the original nonlinear network, Theorem 3 bounds mean and variance errors when the centered remainder satisfies $|I_k(x; \hat{\theta} + \alpha) - \widetilde{I}_k(x; \alpha)| \le B_k \|\alpha\|^2/2$ almost surely with a finite constant B_k. Lemma 5 establishes this condition for a one-hidden-layer ReLU network with nonzero fitted preactivations at the queried inputs, including perturbations that cross activation boundaries. Let a_k stack the layer sensitivities, $v_k = a_k^\top \Sigma_{\tan} a_k$, and $s_k \ge 0$ satisfy $s_k^2 = \frac{B_k^2}{4}[2\,\mathrm{tr}(\Sigma_{\tan}^2) + \mathrm{tr}(\Sigma_{\tan})^2]$, the moment bounds are

$$\begin{aligned} |\mathbb{E}[I_k(x; \hat{\theta} + \alpha)] - \mu_k(x)| &\le \tfrac{B_k}{2}\,\mathrm{tr}(\Sigma_{\tan}), \\ |\,\mathrm{Var}(I_k(x; \hat{\theta} + \alpha)) - v_k| &\le 2\sqrt{v_k}\, s_k + s_k^2. \end{aligned} \tag{12}$$

For the ReLU bound, B_k depends on the activation margins. Lemma 4 separately bounds the effect of curvature approximation on influence covariance. The statements and proofs are in Appendix A; Section 5.5 evaluates propagation accuracy against nonlinear-network samples.

5 Experiments

We test feature-importance recovery (Q1), faithfulness (Q2), uncertainty reliability (Q3), accuracy and efficiency (Q4), regional interpretation (Q5), and multiclass prediction (Q6).

5.1 Experimental Setup

Data and models. ADNI and OASIS test Alzheimer's disease-related structural MRI interpretation; HCP-Aging tests age-group classification with structural MRI and rs-fMRI. We use a 64-ReLU MLP, five-fold subject-level cross-validation, training-only standardization, and last-layer diagonal GGN Laplace (all-layer comparison in Q4). The main ADNI comparison uses CN–AD for its clearer diagnostic contrast; additional tasks include MCI.

Label limitation. Diagnosis and age labels do not specify which ROIs should explain each subject. Regional patterns also vary across patients, so these labels cannot serve as ground truth for feature-importance AUROC, MSE, or R^2.

Metric rationale. On clinical data, we evaluate what can be measured without importance labels: deletion/insertion faithfulness (AOPC/AOPIC), ranking stability (Jacc@10 and Spearman), and anatomical alignment. ROAR tests information removal after retraining; maximum sensitivity tests changes under small input perturbations. Variance agreement and 90% interval coverage assess uncertainty against retraining variability, while HMC provides a posterior-sampling reference. On synthetic data, known feature effects permit direct importance AUROC, MSE, and R^2 evaluation.

Intervention and aggregation. ROI removal measures model reliance under the chosen baseline. Cohort rankings pool $|\mu_k(x)|$ over held-out subjects across folds; AOPC/AOPIC score each subject before averaging. Appendix B defines baselines, implementation, and metrics.

5.2 Q1: Synthetic Experiments

Feature-importance recovery. Friedman #1 and INVASE Syn1–Syn6 provide known feature effects. Table 1 compares CFIP, a five-network ensemble, and SHAP over five seeds, using $n = 2{,}000$ for Friedman. AUROC assesses important-feature identification; MSE and R^2 measure feature-effect recovery. Appendix D.4 gives the full size study.

Benchmark	Method	Instance AUROC	Importance MSE	Importance R^2
Friedman n=2000	CFIP	0.940 ± 0.002	$\mathbf{0.136 \pm 0.019}$	$\mathbf{0.980 \pm 0.003}$
Friedman n=2000	Ensemble	0.948 ± 0.002	0.139 ± 0.008	0.979 ± 0.001
Friedman n=2000	SHAP	0.935 ± 0.003	0.141 ± 0.011	0.977 ± 0.002
INVASE Syn1–Syn6	CFIP	0.816 ± 0.002	0.368 ± 0.007	$\mathbf{0.702 \pm 0.003}$
INVASE Syn1–Syn6	Ensemble	0.819 ± 0.002	0.367 ± 0.005	0.695 ± 0.003
INVASE Syn1–Syn6	SHAP	0.802 ± 0.006	0.386 ± 0.005	0.665 ± 0.004

Table 1: Feature-importance recovery with a SHAP baseline (five seeds, mean $\pm$ SD). INVASE results summarize Syn1–Syn6.

Scalar regression uses the same output-difference construction; a Gaussian likelihood and prior give an exact Gaussian head posterior conditional on the extractor and noise variance (Appendix D.6).

Feature-importance calibration. On INVASE, expected calibration error (ECE) for feature importance is 0.063 ± 0.002 for CFIP and 0.058 ± 0.002 for the five-network ensemble (lower is better).

Depth and extractor uncertainty. Table 2 compares last-layer CFIP with a synthetic variance reference and fifteen refits. The growing gap at depth 10 motivates including extractor uncertainty through the all-layer construction in Section 4.5; Appendix D.5 gives details.

Depth	Parameters	Reference variance	Ensemble variance	CFIP variance
1	2817	$.0161 \pm .0026$	$.0081 \pm .0032$	$\mathbf{.0177 \pm .0028}$
2	7105	$.0221 \pm .0042$	$.0129 \pm .0059$	$\mathbf{.0208 \pm .0026}$
3	11393	$.0255 \pm .0028$	$.0147 \pm .0071$	$\mathbf{.0175 \pm .0027}$
4	15681	$.0254 \pm .0039$	$.0134 \pm .0035$	$\mathbf{.0192 \pm .0023}$
5	19969	$.0312 \pm .0052$	$.0195 \pm .0063$	$\mathbf{.0208 \pm .0043}$
10	41409	$.0330 \pm .0089$	$.0226 \pm .0036$	$.0193 \pm .0028$

Table 2: Depth study; bold CFIP estimates are closer to the reference mean than the ensemble. Last-layer CFIP conditions on the learned extractor; the ensemble refits it.

5.3 Q2: Faithfulness and Stability

On CN–AD, CFIP has the highest mean Jacc@10 and the ensemble the highest mean AOPC (Table 3). ROAR and paired AOPC favor CFIP over IG; Figure 3(a) shows deletion curves. Maximum sensitivity is $\mathbf{0.0087 \pm 0.0006}$ versus 0.0090 ± 0.0006 (Appendix D.2). The influence functional sets the mean ranking; Laplace supplies uncertainty. Global- and cohort-mean interventions preserve CN–AD rankings ($\rho = 1.000$ and 0.993; Appendix B.7).

CFIP also has higher mean AOPC and ranking stability than the strongest comparator for each metric on OASIS and HCP-Aging (Table 4). Appendix B.4 reports every method in this multi-cohort benchmark, including its separate ADNI evaluation.

5.4 Q3: Uncertainty Reliability and Calibration

Variance-rank agreement exceeds 0.90 on all ADNI tasks, but retraining coverage varies: 0.931 on CN–AD, 0.893 on AD–MCI, and 0.708 on CN–MCI (Tables 3 and 18). Against the HMC posterior reference, Laplace preserves variance ordering but underestimates variance and interval coverage (Table 5). Figure 3(b) shows an error signal within confidence bins.

Method	AOPC ↑	Jacc@10 ↑	Spearman ↑	Var-ρ ↑	Cov@90	ROAR AUC ↓
CFIP	0.175 ± 0.020	$\mathbf{0.750 \pm 0.060}$	$\mathbf{0.786 \pm 0.043}$	**0.929**	0.931	0.932 ± 0.012
Ensemble (10)	0.186 ± 0.024	0.740 ± 0.060	0.781 ± 0.043	—	0.917	0.930 ± 0.012
SWAG	0.177 ± 0.021	0.733 ± 0.066	0.785 ± 0.041	0.794	0.929	0.932 ± 0.012
Variational BNN	0.172 ± 0.027	0.737 ± 0.043	0.726 ± 0.022	0.923	0.950	0.934 ± 0.013
IG	0.170 ± 0.025	0.660 ± 0.080	0.710 ± 0.050	—	—	0.935 ± 0.012

Table 3: CN–AD explanation evaluation. Cov@90 targets 0.90; Var-ρ and Cov@90 are point summaries. Evaluation details are in Appendix D.2; other diagnosis tasks are in Table 18. Dashes indicate N/A.

Benchmark	Task	CFIP AOPC ↑	Comparator AOPC ↑	CFIP Stability ↑	Comparator Stability ↑
OASIS	CN vs dementia	$\mathbf{0.148 \pm 0.018}$	0.139 ± 0.020 (IG)	$\mathbf{0.624 \pm 0.031}$	0.604 ± 0.035 (IG)
HCP-Aging	sMRI age group	$\mathbf{0.096 \pm 0.014}$	0.091 ± 0.013 (IG)	$\mathbf{0.588 \pm 0.037}$	0.561 ± 0.034 (IG)
HCP-Aging	rs-fMRI age group	$\mathbf{0.073 \pm 0.012}$	0.070 ± 0.011 (SmoothGrad)	$\mathbf{0.553 \pm 0.044}$	0.527 ± 0.041 (IG)

Table 4: Faithfulness and stability on OASIS and HCP-Aging. Each comparator is the strongest non-CFIP baseline for that metric in the multi-cohort benchmark. Appendix B.4 reports the full results.

A tenfold prior-precision increase preserves variance ordering ($\rho \approx 0.999$) but changes interval width (Appendix D.1). Appendix B.5 gives multi-cohort retraining and sign calibration.

Task	Variance ρ	Median Laplace/HMC variance	HMC-Cov@90
CN–AD	1.000	0.920 ± 0.006	0.868 ± 0.009
AD–MCI	1.000	0.917 ± 0.005	0.871 ± 0.019
CN–MCI	1.000	0.948 ± 0.018	0.831 ± 0.089

Table 5: HMC comparison. HMC-Cov@90 is the fraction of HMC influence draws inside the analytic 90% interval.

5.5 Q4: Pushforward Accuracy and Efficiency

Table 6 compares analytic CFIP moments with draws from the same Laplace posterior. Last-layer and full-tangent evaluations agree closely with their analytic laws. All-parameter draws evaluated in the original nonlinear network give larger variance error and lower coverage, showing the effect of linearization. Appendix B.6 adds logistic-regression, curvature, and cross-cohort checks.

Variant	Model	Variance ρ ↑	RelErr ↓	LapMC-Cov@90
Last-layer	MLP	$\mathbf{0.9998 \pm 0.0000}$	$\mathbf{0.031 \pm 0.014}$	$\mathbf{0.899 \pm 0.008}$
Full tangent	MLP	0.9976 ± 0.0012	0.041 ± 0.018	0.894 ± 0.012
Full nonlinear	MLP	0.943 ± 0.032	0.184 ± 0.071	0.862 ± 0.040

Table 6: Gaussian pushforward validation on ADNI CN–AD. Analytic moments are compared with parameter draws evaluated in the last-layer model, the all-layer affine tangent model, or the original nonlinear network. RelErr measures normalized variance-map error.

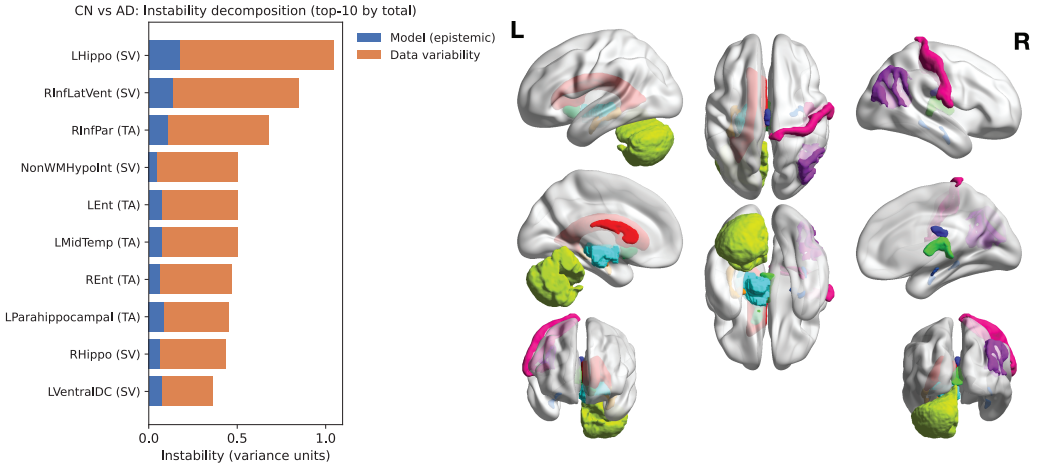
Computational cost. CFIP reuses a precomputed precision across subjects. Table 7 reports explanation time and peak GPU memory for the last-layer path. These measurements exclude training; Appendix D.8 gives the sampling budgets and implementation details.

Method	Precompute (s)	Explain (ms/subject)	Peak GPU (MiB)
CFIP	0.032	**0.056 ± 0.009**	**37.6**
MC-dropout	0.001	0.119 ± 0.036	68.0
BootstrapSHAP	0.467	23.665 ± 0.030	37.8
10-model ensemble	0.006	0.340 ± 0.001	38.5

Table 7: Explanation-phase runtime and peak GPU memory under the stated attribution and sampling configurations. Precompute time excludes model training.

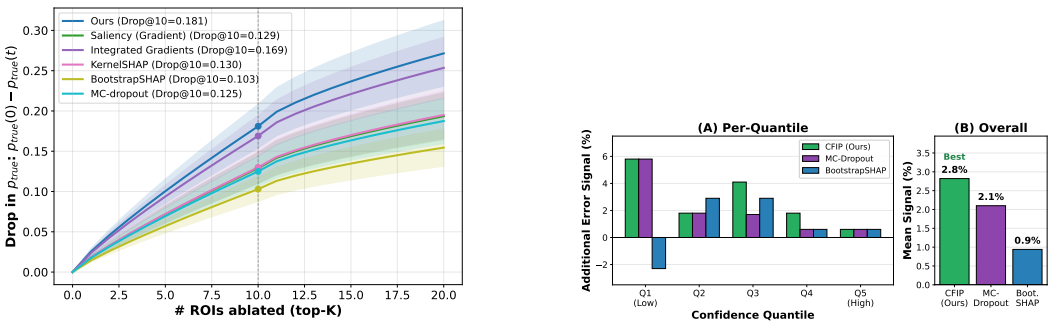
5.6 Q5: Instability Decomposition and Regional Interpretation

Sources of attribution instability. Figure 2(a) decomposes variability as $T_k = M_k + D_k$, where $M_k = \mathbb{E}_x[\Sigma_I(x)_{kk}]$ measures model uncertainty and $D_k = \mathrm{Var}_x(\mu_k(x))$ measures data heterogeneity. Data heterogeneity dominates for most top CN–AD regions: variation in mean regional effects across subjects exceeds parameter-induced uncertainty.



(a) Instability decomposition. (b) Top influential ROIs.

Figure 2: ADNI CN vs. AD regional interpretation. (a) Model uncertainty and data heterogeneity for ROIs ranked by total attribution variability. (b) Top ten ROIs ranked by subject-average absolute posterior mean.



(a) Deletion faithfulness on ADNI CN vs. AD. (b) Error signal within confidence bins.

Figure 3: Faithfulness and uncertainty on CN–AD. (a) Higher deletion curves indicate more faithful ROI rankings. (b) Within each confidence quintile, the error-rate gap between high- and low-uncertainty groups measures the additional error signal; larger positive gaps indicate better separation.

When uncertainty changes interpretation. Across ten retrainings of CN–AD fold 0, fourteen ROIs enter the top-ten lists: six appear in every fit and eight intermittently. The intermittent group has 1.24 times the uncertainty and a sign-resolution fraction of 72%, versus 92% for the always-selected group. ST52TA appears in only one of ten lists, with sign resolved in 51% of subjects. A single importance map could present it as a leading feature; CFIP attaches a reliability signal that calls for a more cautious regional interpretation (Appendix D.7).

Regional interpretation. Figure 2(b) maps the ten most influential ROIs. We compare them with an AD-related set covering hippocampal, entorhinal, parahippocampal, amygdala, middle temporal, and ventricular regions (Braak & Braak, 1991; Jack et al., 2010; Rathore et al., 2017; Cho et al., 2012). Six of the top ten regions overlap on ADNI CN–AD and five on OASIS CN–dementia. Table 8 gives anatomical alignment and sign confidence across tasks. These results describe the model's regional reliance under ROI removal, with anatomical overlap providing a separate plausibility check.

Benchmark	Top-10 overlap	Enrichment p	Mean sign conf.
ADNI CN vs. AD	6 / 10	0.006	0.94 ± 0.03
OASIS CN vs. dementia	5 / 10	0.019	0.91 ± 0.04
ADNI AD vs. MCI	5 / 10	0.031	0.88 ± 0.05
ADNI CN vs. MCI	4 / 10	0.074	0.84 ± 0.06

Table 8: Neuroanatomical alignment. Top-10 overlap counts ROIs in the AD-related set described in the text. Enrichment uses Fisher's exact test; the p-values provide exploratory evidence of anatomical alignment.

5.7 Q6: Multiclass Extension

For CN/MCI/AD classification, CFIP uses the class-logit difference $I_{kc}(x) = f_c(x) - f_c(x^{(-k)})$, which is affine in the class-c output weights given a fixed extractor. Table 9 shows AOPC close to the ensemble and higher mean stability than IG and BootstrapSHAP.

Method	Deletion AOPC	Ranking stability ρ
CFIP	$.0819 \pm .0126$	$.751 \pm .072$
Integrated Gradients	$.0774 \pm .0127$	$.504 \pm .073$
BootstrapSHAP	$.0621 \pm .0174$	$.603 \pm .066$
10-model ensemble	$.0820 \pm .0206$	$.773 \pm .037$

Table 9: Deletion faithfulness and ranking stability on three-class CN/MCI/AD classification. The ensemble has the highest mean scores.

6 Conclusion

We presented CFIP, a reliability layer that turns ROI interventions and Laplace-linearized uncertainty into a closed-form posterior over regional influences. Across structural MRI and fMRI benchmarks, CFIP provides faithful regional explanations and uncertainty estimates that track observed attribution variability. Its analytic influence posterior supports sampling-free inference and separates model uncertainty from data heterogeneity, giving a distributional view of neuroimaging model explanations.

AI USAGE STATEMENT

During preparation of this revision, AI assistance was used to organize and edit this manuscript, transcribe supplied result tables, inspect mathematical claims, check citation context, and review presentation. The assistance did not generate the experimental measurements reported here.

REPRODUCIBILITY STATEMENT

Code is available at `https://anonymous.4open.science/r/CFIP-7BE1/`. Section 4 and Appendix A state the modeling assumptions and provide the derivations and proofs; Appendix C details the full-network tangent construction. Section 5 and Appendix B describe the datasets, preprocessing, model settings, and evaluation metrics. Appendix D documents the additional validation studies, including synthetic benchmarks and posterior checks.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Hirotaka Akita, Kosuke Nakago, Tomoki Komatsu, Yohei Sugawara, Shin-ichi Maeda, Yukino Baba, and Hisashi Kashima. Bayesgrad: Explaining predictions of graph convolutional networks. In *International Conference on Neural Information Processing*, pp. 81–92. Springer, 2018.

Bharat Biswal, F. Zerrin Yetkin, Victor M. Haughton, and James S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995. doi: 10.1002/mrm.1910340409.

Susan Y. Bookheimer, David H. Salat, Melissa Terpstra, Beau M. Ances, Deanna M. Barch, Randy L. Buckner, Gregory C. Burgess, Sandra W. Curtiss, Mirella Diaz-Santos, Jennifer Stine Elam, et al. The lifespan human connectome project in aging: An overview. *NeuroImage*, 185:335–348, 2019. doi: 10.1016/j.neuroimage.2018.10.009.

Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.

Youngsang Cho, Joon Kyung Seong, Yong Jeong, and Sung Yong Shin. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. *NeuroImage*, 59(3):2217–2230, February 2012. doi: 10.1016/j.neuroimage.2011.09.085. URL https://doi.org/10.1016/j.neuroimage.2011.09.085.

Santiago Cifuentes, Leopoldo E. Bertossi, Nina Pardal, Sergio Abriola, Maria Vanina Martinez, and Miguel Romero. The distributional uncertainty of the SHAP score in explainable machine learning. In *ECAI 2024 - 27th European Conference on Artificial Intelligence (including PAIS 2024)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pp. 971–978. IOS Press, 2024. doi: 10.3233/FAIA240586. URL https://doi.org/10.3233/FAIA240586.

Rémi Cuingnet et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage*, 56(2):766–781, 2011. doi: 10.1016/j.neuroimage.2010.06.013.

Erik A. Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20089–20103, 2021. URL https://arxiv.org/abs/2106.14806.

Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006. doi: 10.1016/j.neuroimage.2006.01.021.

Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1050–1059. PMLR, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019. doi: 10.1609/aaai.v33i01.33013681.

Michael P. Harms, Leah H. Somerville, Beau M. Ances, Jesper Andersson, Deanna M. Barch, Matteo Bastiani, Susan Y. Bookheimer, Timothy B. Brown, Randy L. Buckner, Gregory C. Burgess, et al. Extending the Human Connectome Project across ages: Imaging protocols for the lifespan development and aging projects. *NeuroImage*, 183:972–984, 2018. doi: 10.1016/j.neuroimage.2018.09.060.

Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

Clifford R. Jr. Jack et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008. doi: 10.1002/jmri.21049.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8580–8589, 2018. URL https://papers.nips.cc/paper/8076-neural-tangent-kernel-convergence-and-generalization-in-neural-networks.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of saliency methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pp. 267–280. Springer, 2019. doi: 10.1007/978-3-030-28954-6_14.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Pamela J. LaMontagne, Tammie L. S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv*, 2019. doi: 10.1101/2019.12.13.19014902. URL https://www.medrxiv.org/content/10.1101/2019.12.13.19014902v1.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8570–8581, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0d1a9651497a38d8b1c3871c84528bd4-Abstract.html.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992. doi: 10.1162/neco.1992.4.3.415.

Ronald C. Petersen, Paul S. Aisen, Laurel A. Beckett, Michael C. Donohue, Anthony C. Gamst, Danielle J. Harvey, Clifford R. Jr. Jack, William J. Jagust, Leslie M. Shaw, Arthur W. Toga, John Q. Trojanowski, and Michael W. Weiner. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology*, 74(3):201–209, 2010. doi: 10.1212/WNL.0b013e3181cb3e25.

Jonathan D. Power, Alexander L. Cohen, Steven M. Nelson, Gagan S. Wig, Kelly Anne Barnes, Jessica A. Church, Alecia C. Vogel, Timothy O. Laumann, Fran M. Miezin, Bradley L. Schlaggar, and Steven E. Petersen. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011. doi: 10.1016/j.neuron.2011.09.006.

Saima Rathore, Mohamad Habes, M. Aksam Iftikhar, Allison Shacklett, and Christos Davatzikos. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, 155:530–548, 2017. doi: 10.1016/j.neuroimage.2017.03.057.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Skdvd2xAZ.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations Workshop*, 2014. URL https://arxiv.org/abs/1312.6034.

Stephen M. Smith, Christian F. Beckmann, Jesper Andersson, Edward J. Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A. Feinberg, Ludovica Griffanti, Michael P. Harms, et al. Resting-state fMRI in the Human Connectome Project. *NeuroImage*, 80:144–168, 2013. doi: 10.1016/j.neuroimage.2013.05.039.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3319–3328. PMLR, 2017.

Martijn P. van den Heuvel and Hilleke E. Hulshoff Pol. Exploring the brain network: a review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, 2010. doi: 10.1016/j.euroneuro.2010.03.008.

A Proofs for Approximation Analysis

Proof of Theorem 1. For each ROI k, using the assumed linear stochastic parameterization,

$$I_k(x;\theta) = f_\theta(x) - f_\theta(x^{(-k)}) = \langle \theta, \phi(x)\rangle - \langle \theta, \phi(x^{(-k)})\rangle = \theta^\top\big(\phi(x) - \phi(x^{(-k)})\big).$$

By definition, $A_{k,:}(x) = (\phi(x) - \phi(x^{(-k)}))^\top$, hence

$$I_k(x;\theta) = A_{k,:}(x)\theta.$$

Stacking the m ROI influences gives

$$I(x;\theta) = A(x)\theta.$$

Therefore $I(x;\theta)$ is affine, in fact linear, in θ, and its Jacobian with respect to θ is exactly

$$J_I(x) = A(x).$$

Thus the first-order linearization in Eq. (6) has zero remainder.

Now let $\theta \sim \mathcal{N}(\hat{\theta}, \Sigma)$. Since a linear transformation of a Gaussian random vector is Gaussian,

$$A(x)\theta \sim \mathcal{N}\Big(A(x)\hat{\theta},\ A(x)\Sigma A(x)^\top\Big).$$

Because $I(x;\theta) = A(x)\theta$, the induced influence distribution is

$$I(x) \mid \mathcal{D} \sim \mathcal{N}\Big(A(x)\hat{\theta},\ A(x)\Sigma A(x)^\top\Big).$$

This proves the result. Logistic regression is obtained by taking $\phi(x) = x$. Last-layer Laplace models are also covered conditional on the fixed feature extractor, with $\phi(x)$ equal to the last-layer feature vector. □

Proof of Proposition 2. Write the perturbation as $\alpha = (\alpha_h, \alpha_e)$ and let $\alpha \sim \mathcal{N}(0, H^{-1})$. Since $H \succ 0$, both C and its Schur complement $S = D - B^\top C^{-1} B$ are positive definite. Completing the square,

$$\alpha^\top H \alpha = (\alpha_h + C^{-1}B\alpha_e)^\top C(\alpha_h + C^{-1}B\alpha_e) + \alpha_e^\top S \alpha_e.$$

Consequently $\epsilon = \alpha_h + C^{-1}B\alpha_e$ and α_e are independent, with $\epsilon \sim \mathcal{N}(0, C^{-1})$ and $\alpha_e \sim \mathcal{N}(0, S^{-1})$. Thus

$$J\alpha = J_h\epsilon + (J_e - J_h C^{-1}B)\alpha_e,$$

and independence gives Eq. (11). The second term is positive semidefinite because $S^{-1} \succ 0$. Conditioning on $\alpha_e = 0$ gives head covariance C^{-1}, which identifies the first term. This is the Gaussian conditional law, not the head's marginal covariance. The additional contribution can vanish in directions for which G is zero. With block-diagonal H, it reduces to $J_e D^{-1} J_e^\top$. The comparison uses a common precision; it does not order separately chosen curvature approximations. □

Theorem 3 (Full-network tangent law and nonlinear moment error)**.** *Fix $\hat{\theta}$ and inputs $x, x^{(-1)}, \ldots, x^{(-m)}$ at which the logit is differentiable in its parameters. Let $a_k = \psi_{\hat{\theta}}(x) - \psi_{\hat{\theta}}(x^{(-k)})$, stack $a_k^\top$ in $A_{\tan}$, and take $\alpha \sim \mathcal{N}(0, \Sigma_{\tan})$. Then the tangent influence satisfies*

$$\widetilde{I}(x;\alpha) = I(x;\hat{\theta}) + A_{\tan}\alpha \sim \mathcal{N}\Big(I(x;\hat{\theta}), A_{\tan}\Sigma_{\tan}A_{\tan}^\top\Big).$$

For the nonlinear moment bounds, suppose separately that $R_k(\alpha) = I_k(x;\hat{\theta}+\alpha) - \widetilde{I}_k(x;\alpha)$ satisfies $|R_k(\alpha)| \le B_k\|\alpha\|_2^2/2$ almost surely with one finite $B_k \ge 0$ independent of α. Define v_k and $s_k \ge 0$ by

$$v_k = a_k^\top \Sigma_{\tan} a_k, \qquad s_k^2 = \frac{B_k^2}{4}\left[2\,\mathrm{tr}(\Sigma_{\tan}^2) + \mathrm{tr}(\Sigma_{\tan})^2\right].$$

Then

$$\Big|\mathbb{E}[I_k(x;\hat{\theta}+\alpha)] - I_k(x;\hat{\theta})\Big| \le \frac{B_k}{2}\,\mathrm{tr}(\Sigma_{\tan}),$$
$$\Big|\mathrm{Var}(I_k(x;\hat{\theta}+\alpha)) - v_k\Big| \le 2\sqrt{v_k}\,s_k + s_k^2.$$

Proof. The first claim follows from the affine Gaussian transformation rule and requires no remainder bound. For the moment statements, write $I_k(x;\hat{\theta}+\alpha) = \mu_k + L_k + R_k$, where $\mu_k = I_k(x;\hat{\theta})$ and $L_k = a_k^\top \alpha$. Diagonalizing the positive semidefinite covariance gives $\|\alpha\|^2 = \sum_j \gamma_j Z_j^2$, where Z_j are independent standard Normals and $\gamma_j \geq 0$ are its eigenvalues. Since $\mathbb{E}Z_j^2 = 1$, $\mathbb{E}Z_j^4 = 3$, and $\mathbb{E}Z_i^2 Z_j^2 = 1$ for $i \neq j$,

$$\mathbb{E}\|\alpha\|^2 = \operatorname{tr}(\Sigma_{\tan}), \qquad \mathbb{E}\|\alpha\|^4 = 2\operatorname{tr}(\Sigma_{\tan}^2) + \operatorname{tr}(\Sigma_{\tan})^2.$$

The remainder hypothesis therefore ensures $R_k \in L^2$ and yields

$$|\mathbb{E}R_k| \leq \frac{B_k}{2}\operatorname{tr}(\Sigma_{\tan}), \qquad \operatorname{Var}(R_k) \leq \mathbb{E}R_k^2 \leq s_k^2.$$

Because $\mathbb{E}L_k = 0$, the first inequality proves the mean bound. For the variance, expansion and Cauchy–Schwarz give

$$\begin{aligned} |\operatorname{Var}(L_k + R_k) - v_k| &= |\operatorname{Var}(R_k) + 2\operatorname{Cov}(L_k, R_k)| \\ &\leq \operatorname{Var}(R_k) + 2\sqrt{v_k \operatorname{Var}(R_k)} \leq s_k^2 + 2\sqrt{v_k}\, s_k. \end{aligned}$$

The same argument includes singular covariance and $v_k = 0$. □

A sufficient smoothness condition. For a fixed perturbation α, suppose $f_\theta(z)$ is C^2 on a neighborhood of the segment $\hat{\theta} + t\alpha$, $0 \leq t \leq 1$, and its parameter Hessian has spectral norm at most B_z there. The integral Taylor remainder satisfies

$$R_z(\alpha) = \int_0^1 (1-t)\alpha^\top \nabla_\theta^2 f_{\hat{\theta}+t\alpha}(z)\alpha\, dt, \qquad |R_z(\alpha)| \leq \frac{B_z}{2}\|\alpha\|^2.$$

Subtracting the remainders at x and $x^{(-k)}$ gives the influence bound with $B_k = B_x + B_{x^{(-k)}}$. The unconditional moment result requires one finite B_k for almost every Gaussian perturbation; a global C^2 logit with uniformly bounded Hessian is sufficient. The ReLU result below supplies a different sufficient condition directly on the centered remainder.

Lemma 4 (Curvature-error propagation)**.** *Let*

$$\Sigma_I(x) = J_I(x)\Sigma J_I(x)^\top, \qquad \tilde{\Sigma}_I(x) = J_I(x)\tilde{\Sigma} J_I(x)^\top.$$

Then

$$\tilde{\Sigma}_I(x) - \Sigma_I(x) = J_I(x)(\tilde{\Sigma} - \Sigma)J_I(x)^\top, \qquad \|\tilde{\Sigma}_I(x) - \Sigma_I(x)\|_2 \leq \|J_I(x)\|_2^2 \|\tilde{\Sigma} - \Sigma\|_2.$$

For row $j_k(x)^\top$ of $J_I(x)$,

$$\left|[\tilde{\Sigma}_I(x)]_{kk} - [\Sigma_I(x)]_{kk}\right| \leq \|j_k(x)\|_2^2 \|\tilde{\Sigma} - \Sigma\|_2.$$

If $\Sigma = H^{-1}$ and $\tilde{\Sigma} = \tilde{H}^{-1}$ for positive definite $H, \tilde{H}$, then

$$\|\tilde{\Sigma} - \Sigma\|_2 \leq \|H^{-1}\|_2 \, \|H - \tilde{H}\|_2 \, \|\tilde{H}^{-1}\|_2.$$

Proof of Lemma 4. By definition,

$$\Sigma_I(x) = J_I(x)\Sigma J_I(x)^\top, \qquad \tilde{\Sigma}_I(x) = J_I(x)\tilde{\Sigma} J_I(x)^\top.$$

Subtracting gives

$$\tilde{\Sigma}_I(x) - \Sigma_I(x) = J_I(x)\tilde{\Sigma} J_I(x)^\top - J_I(x)\Sigma J_I(x)^\top = J_I(x)(\tilde{\Sigma} - \Sigma)J_I(x)^\top.$$

Taking spectral norms and using submultiplicativity,

$$\|\tilde{\Sigma}_I(x) - \Sigma_I(x)\|_2 = \|J_I(x)(\tilde{\Sigma} - \Sigma)J_I(x)^\top\|_2 \leq \|J_I(x)\|_2 \|\tilde{\Sigma} - \Sigma\|_2 \|J_I(x)^\top\|_2.$$

Since $\|J_I(x)^\top\|_2 = \|J_I(x)\|_2$,

$$\|\tilde{\Sigma}_I(x) - \Sigma_I(x)\|_2 \leq \|J_I(x)\|_2^2 \|\tilde{\Sigma} - \Sigma\|_2.$$

For the diagonal entry corresponding to ROI k, let $j_k(x)^\top$ be the k-th row of $J_I(x)$. Then

$$[\tilde{\Sigma}_I(x)]_{kk} - [\Sigma_I(x)]_{kk} = j_k(x)^\top(\tilde{\Sigma} - \Sigma)j_k(x).$$

Using $|u^\top Bu| \le \|B\|_2\|u\|_2^2$ gives

$$\left|[\tilde{\Sigma}_I(x)]_{kk} - [\Sigma_I(x)]_{kk}\right| \le \|j_k(x)\|_2^2\|\tilde{\Sigma} - \Sigma\|_2.$$

Finally, suppose $\Sigma = H^{-1}$ and $\tilde{\Sigma} = \tilde{H}^{-1}$ for positive definite H and $\tilde{H}$. Then

$$\tilde{\Sigma} - \Sigma = \tilde{H}^{-1} - H^{-1} = H^{-1}(H - \tilde{H})\tilde{H}^{-1}.$$

Taking spectral norms yields

$$\|\tilde{\Sigma} - \Sigma\|_2 \le \|H^{-1}\|_2\|H - \tilde{H}\|_2\|\tilde{H}^{-1}\|_2.$$

This proves the lemma. □

Lemma 5 (Centered remainder for a shallow ReLU network)**.** *Let $f_\theta(z) = \sum_{j=1}^h v_j\rho(u_j^\top \bar{z}) + c$, where $\rho(t) = \max(t, 0)$, $\bar{z} = (z, 1)$ includes the hidden bias coordinate, and $\theta = (U, v, c)$ contains all parameters. Assume $t_j(z) = \hat{u}_j^\top \bar{z} \neq 0$ for every unit and each queried input $z \in \{x, x^{(-1)}, \dots, x^{(-m)}\}$. Set*

$$C_z = \|\bar{z}\|^2 \max_j \frac{|\hat{v}_j|}{4|t_j(z)|}, \qquad B_z = 2C_z + \|\bar{z}\|.$$

For every perturbation α, the centered logit remainder obeys

$$|f_{\hat{\theta}+\alpha}(z) - f_{\hat{\theta}}(z) - \nabla_\theta f_{\hat{\theta}}(z)^\top\alpha| \le \frac{B_z}{2}\|\alpha\|^2.$$

Hence $B_k = B_x + B_{x^{(-k)}}$ satisfies the remainder hypothesis of Theorem 3 for the ROI influence.

Proof. For $t \neq 0$, define $r(t, d) = \rho(t + d) - \rho(t) - \mathbf{1}_{t>0}d$. Without a gate crossing, $r = 0$. At a crossing, $|d| \ge |t|$ and $|r| = |d| - |t|$. The inequality $4|t|(|d| - |t|) \le d^2$, equivalent to $(|d| - 2|t|)^2 \ge 0$, proves

$$|r(t, d)| \le \frac{d^2}{4|t|}.$$

Write $\alpha = (\Delta U, \Delta v, \Delta c)$ and $d_j = \Delta u_j^\top \bar{z}$. Expanding the network output and subtracting its parameter tangent gives

$$R_z = \sum_j \hat{v}_j r(t_j, d_j) + \sum_j \Delta v_j\big[\rho(t_j + d_j) - \rho(t_j)\big].$$

The bias Δc cancels. The scalar inequality and $\sum_j d_j^2 \le \|\bar{z}\|^2\|\Delta U\|_F^2$ bound the first sum. The ReLU Lipschitz property and Cauchy–Schwarz bound the second, giving

$$\begin{aligned}|R_z| &\le C_z\|\Delta U\|_F^2 + \|\bar{z}\|\|\Delta v\|\|\Delta U\|_F \\ &\le \left(C_z + \frac{\|\bar{z}\|}{2}\right)\|\alpha\|^2.\end{aligned}$$

The triangle inequality for the two input remainders proves the influence bound. This argument permits activation crossings. The constant is finite under the stated margin condition and can be large near a zero preactivation. □

The shallow-network lemma gives a sufficient condition for the architecture used in the main experiments. Its application to a trained model requires checking the preactivation margins. For deeper networks, a fixed activation pattern can still give higher-order dependence on the parameters, so this particular global quadratic bound is not automatic. The tangent Gaussian law continues to hold whenever the fitted gradients exist.

Benchmark	Modality	Task	#Subj.	Acc.	F1	AUC
ADNI	sMRI ROI	CN vs AD	1718	91.2 ± 1.3	85.0 ± 2.2	0.947 ± 0.010
ADNI	sMRI ROI	AD vs MCI	1930	74.0 ± 3.6	80.8 ± 3.5	0.806 ± 0.013
ADNI	sMRI ROI	CN vs MCI	2648	64.5 ± 2.4	65.1 ± 2.2	0.716 ± 0.031
OASIS	sMRI ROI	CN vs dementia	1042	84.6 ± 2.1	78.8 ± 2.9	0.902 ± 0.018
HCP-Aging	sMRI ROI	age group	672	77.9 ± 2.5	77.4 ± 2.8	0.854 ± 0.022
HCP-Aging	rs-fMRI FC	age group	612	72.4 ± 3.0	71.9 ± 3.1	0.789 ± 0.030

Table 10: Predictive performance of the MLP models in the multi-cohort benchmark. Results are mean $\pm$ standard deviation over 5-fold subject-level cross-validation. ADNI uses 118 FreeSurfer FS7 ROI features from the earliest diagnosis-consistent visit. Non-ADNI benchmarks evaluate whether CFIP transfers beyond a single ADNI structural-ROI pipeline.

B Additional Experimental Details

B.1 Multi-Cohort Benchmark Predictive Performance

B.2 Implementation Details

Setup. We evaluate CFIP on multi-cohort regional neuroimaging benchmarks. ADNI and OASIS test Alzheimer's disease-related structural MRI interpretation, while HCP-Aging tests age-group classification using structural MRI and rs-fMRI connectivity. Structural MRI ROIs use anatomical measurements such as cortical thickness or volume; for connectivity, intervening on ROI k replaces edges incident to that region. Appendix B.1 reports predictive performance for the multi-cohort benchmark.

Models. The benchmark predictor is an MLP with 64 ReLU hidden units, five-fold subject-level cross-validation, and training-only standardization. CFIP uses last-layer diagonal GGN Laplace unless otherwise stated. Q4 also evaluates full-network tangent CFIP and nonlinear full-network sampling on ADNI CN vs. AD.

Baselines. We compare CFIP with a ten-model deep ensemble, SWAG, and a variational BNN using Bayes-by-Backprop. The multi-cohort benchmark includes Gradient, SmoothGrad, Integrated Gradients (IG), KernelSHAP, BootstrapSHAP, and MC-dropout. We also evaluate CFIP-Grad, which uses the same Laplace posterior with a gradient-based influence functional, and Ensemble-Ablation, which computes ROI intervention effects across independently retrained models.

Metrics. We evaluate faithfulness with deletion AOPC and insertion AOPIC, and ranking stability with Jacc@10 and Spearman correlation. For uncertainty, we measure variance-rank agreement and empirical coverage of 90% influence intervals (Cov@90). Retraining variability provides an empirical reference for explanation reliability; HMC samples assess the posterior approximation; Monte Carlo Laplace samples assess the analytic pushforward. The appendix provides full metric definitions and sign-confidence results.

Component	Setting
Splits	5-fold subject-level cross-validation
Standardization	training split only, then applied to validation/test splits
Main predictor	one-hidden-layer MLP with 64 ReLU hidden units
Linear sanity check	logistic regression with ℓ_2 regularization
CFIP posterior	last-layer diagonal GGN Laplace unless otherwise stated
Ensemble proxy	independently retrained MLPs with different random seeds
IG	path-integrated gradients with fixed baseline
SmoothGrad	gradient saliency averaged over noisy inputs
MC-dropout	stochastic test-time dropout attribution samples
KernelSHAP	perturbation-based SHAP approximation with fixed background set
BootstrapSHAP	SHAP estimates over bootstrapped background sets

Table 11: Implementation details for the multi-cohort interpretation benchmark. The ensemble provides an empirical reference for retraining variability.

B.3 METRIC DEFINITIONS

Let $p_y(x)$ denote the model probability assigned to the true class y, and let $s(x) \in \mathbb{R}^m$ be an attribution score vector. ROIs are ranked by decreasing $|s_k(x)|$. Let S_K be the set of top-K ROIs, and let $x^{(-S_K)}$ denote the input obtained by replacing ROIs in S_K by their baselines.

Deletion faithfulness. For deletion, top-ranked ROIs are progressively removed. We define

$$\mathrm{AOPC} = \frac{1}{K_{\max}} \sum_{K=1}^{K_{\max}} \left[p_y(x) - p_y(x^{(-S_K)}) \right],$$

and

$$\mathrm{Drop@10} = p_y(x) - p_y(x^{(-S_{10})}).$$

Larger values indicate that the selected ROIs more strongly support the model's original prediction.

Insertion faithfulness. For insertion, we start from a baseline input $x^{(0)}$ and progressively restore top-ranked ROIs. Let $x^{(+S_K)}$ denote the input where only ROIs in S_K are restored to their original values. We define

$$\mathrm{AOPIC} = \frac{1}{K_{\max}} \sum_{K=1}^{K_{\max}} \left[p_y(x^{(+S_K)}) - p_y(x^{(0)}) \right].$$

AOPIC measures confidence recovery and therefore provides a non-deletion sufficiency check.

Stability. For each retrain or fold, we compute a global ROI ranking by averaging $|s_k(x)|$ over test subjects. Stability is the average Spearman correlation between global ROI rankings across retrains.

Attribution reference. The uncertainty scores evaluate each method's attribution functional together with its uncertainty estimator. Coverage and sign comparisons therefore measure the performance of the complete explanation pipeline.

Variance agreement. For uncertainty-aware methods, let V_{ik}^{method} denote the estimated attribution variance for subject i and ROI k, and let V_{ik}^{ens} denote the empirical variance across retrained ensemble models. We report Spearman correlation over all subject–ROI pairs:

$$\mathrm{VarCorr\text{-}S} = \rho_{\mathrm{Spearman}} \left(\{V_{ik}^{\mathrm{method}}\}_{i,k}, \{V_{ik}^{\mathrm{ens}}\}_{i,k} \right).$$

We also report the normalized relative error

$$\mathrm{VarRelErr} = \frac{\left\| \bar{V}^{\mathrm{method}} - \bar{V}^{\mathrm{ens}} \right\|_1}{\left\| \bar{V}^{\mathrm{ens}} \right\|_1},$$

where $\bar{V}$ denotes variance maps normalized to sum to one. This measures variance-map shape, not absolute variance scale.

Coverage. For CFIP, the marginal posterior for subject i and ROI k is

$$I_{ik} \mid \mathcal{D} \sim \mathcal{N}(\mu_{ik}, \sigma_{ik}^2), \qquad \sigma_{ik}^2 = [\Sigma_I(x_i)]_{kk}.$$

The central 90% interval is

$$\mathcal{C}_{ik}^{90} = [\mu_{ik} - 1.645\sigma_{ik}, \mu_{ik} + 1.645\sigma_{ik}].$$

For sampling-based baselines, the 90% interval is the empirical 5–95% quantile interval over attribution samples.

For ensemble coverage, with retrained-model influences $I_{ik}^{(e)}$,

$$\mathrm{Ens\text{-}Cov@90} = \frac{1}{nmE} \sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{e=1}^{E} \mathbf{1}\{I_{ik}^{(e)} \in \mathcal{C}_{ik}^{90}\}.$$

The coverage error is

$$\mathrm{CovErr} = |\mathrm{Ens\text{-}Cov@90} - 0.90|.$$

Sign calibration. CFIP's analytic sign confidence is

$$p_{ik}^{\mathrm{CFIP}} = \mathbb{P}(I_{ik} > 0 \mid \mathcal{D}) = \Phi\left(\frac{\mu_{ik}}{\sigma_{ik}}\right).$$

This CDF expression applies for $\sigma_{ik} > 0$; at zero variance use $\mathbf{1}\{\mu_{ik} > 0\}$. The empirical sign frequency under retraining is

$$p_{ik}^{\mathrm{ens}} = \frac{1}{E}\sum_{e=1}^{E} \mathbf{1}\{I_{ik}^{(e)} > 0\}.$$

We report the squared discrepancy between sign probabilities and retraining frequencies

$$\text{Sign-Brier} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m}\left(p_{ik}^{\mathrm{CFIP}} - p_{ik}^{\mathrm{ens}}\right)^2.$$

Laplace–MC pushforward validation. To validate the analytic pushforward, we draw $\theta^{(s)} \sim \mathcal{N}(\hat{\theta}, \Sigma)$ and recompute $I_k(x_i; \theta^{(s)})$. LapMC Spearman is the rank correlation between analytic CFIP variances and Monte Carlo variances over subject–ROI pairs. LapMC-Cov@90 is the fraction of Monte Carlo influence samples falling inside the analytic CFIP 90% intervals.

B.4 Multi-Cohort Faithfulness and Stability Results

This benchmark compares CFIP with attribution baselines on ADNI, OASIS, and HCP-Aging. Its ADNI evaluation is separate from the Bayesian comparison in Table 3.

See Table 12.

B.5 Full Uncertainty Reliability and Calibration Results

We evaluate uncertainty against retraining variability for the multi-cohort benchmark above.

See Table 13.

B.6 Full Validation Results

See Table 14, Table 15 and Table 8.

B.7 Baseline Sensitivity

See Table 16.

B.8 Additional Error-Signal Analysis

Samples are first grouped by prediction confidence and then split within each confidence bin into high-uncertainty and low-uncertainty groups. A positive error gap indicates that the uncertainty estimate identifies samples more likely to be misclassified even within confidence bins. See Figure 3b.

C Full-Network Tangent Construction

CFIP can be instantiated with different stochastic parameter blocks. Our main experiments use last-layer Laplace because the feature extractor is fixed and the logit is affine in the stochastic parameters, making the Gaussian pushforward exact conditional on the Laplace posterior. To incorporate uncertainty from all layers, we also consider a full-network tangent extension.

Let $\hat{\theta}$ be the trained network and define the tangent feature

$$\psi_{\hat{\theta}}(z) = \nabla_\theta f_\theta(z)\big|_{\theta=\hat{\theta}}.$$

Benchmark	Task	Method	AOPC ↑	Drop@10 ↑	Stability-S ↑
ADNI	CN vs. AD	CFIP	**0.172 ± 0.020**	**0.181 ± 0.028**	**0.652 ± 0.025**
		Gradient	0.127 ± 0.016	0.129 ± 0.023	0.616 ± 0.024
		SmoothGrad	0.130 ± 0.017	0.132 ± 0.023	0.620 ± 0.025
		Integrated Gradients	0.162 ± 0.019	0.169 ± 0.026	0.643 ± 0.029
		KernelSHAP	0.096 ± 0.024	0.130 ± 0.033	0.581 ± 0.020
		BootstrapSHAP	0.117 ± 0.011	0.103 ± 0.015	0.606 ± 0.022
		MC-dropout	0.124 ± 0.014	0.125 ± 0.022	0.617 ± 0.023
ADNI	AD vs. MCI	CFIP	**0.064 ± 0.013**	**0.066 ± 0.014**	**0.611 ± 0.059**
		Gradient	0.048 ± 0.004	0.047 ± 0.003	0.584 ± 0.060
		SmoothGrad	0.049 ± 0.005	0.048 ± 0.004	0.586 ± 0.058
		Integrated Gradients	0.063 ± 0.013	0.065 ± 0.014	0.605 ± 0.060
		KernelSHAP	0.033 ± 0.012	0.042 ± 0.016	0.567 ± 0.063
		BootstrapSHAP	0.044 ± 0.007	0.040 ± 0.007	0.596 ± 0.062
		MC-dropout	0.047 ± 0.004	0.046 ± 0.003	0.584 ± 0.060
ADNI	CN vs. MCI	CFIP	0.081 ± 0.026	0.089 ± 0.030	**0.541 ± 0.022**
		Gradient	0.064 ± 0.025	0.068 ± 0.034	0.506 ± 0.017
		SmoothGrad	0.066 ± 0.026	0.070 ± 0.033	0.510 ± 0.018
		Integrated Gradients	0.080 ± 0.025	0.085 ± 0.028	0.530 ± 0.019
		KernelSHAP	0.074 ± 0.028	**0.093 ± 0.034**	0.495 ± 0.038
		BootstrapSHAP	**0.089 ± 0.030**	0.091 ± 0.035	0.509 ± 0.045
		MC-dropout	0.063 ± 0.025	0.066 ± 0.030	0.506 ± 0.019
OASIS	CN vs. dementia	CFIP	**0.148 ± 0.018**	**0.155 ± 0.023**	**0.624 ± 0.031**
		Gradient	0.112 ± 0.017	0.118 ± 0.021	0.582 ± 0.033
		SmoothGrad	0.116 ± 0.018	0.122 ± 0.022	0.587 ± 0.032
		Integrated Gradients	0.139 ± 0.020	0.147 ± 0.023	0.604 ± 0.035
		KernelSHAP	0.090 ± 0.021	0.104 ± 0.027	0.552 ± 0.040
		BootstrapSHAP	0.107 ± 0.019	0.112 ± 0.025	0.570 ± 0.038
		MC-dropout	0.110 ± 0.017	0.116 ± 0.022	0.578 ± 0.034
HCP-Aging	sMRI age group	CFIP	**0.096 ± 0.014**	**0.101 ± 0.018**	**0.588 ± 0.037**
		Gradient	0.071 ± 0.012	0.076 ± 0.016	0.535 ± 0.035
		SmoothGrad	0.084 ± 0.013	0.089 ± 0.017	0.552 ± 0.036
		Integrated Gradients	0.091 ± 0.013	0.096 ± 0.018	0.561 ± 0.034
		KernelSHAP	0.061 ± 0.015	0.070 ± 0.019	0.510 ± 0.042
		BootstrapSHAP	0.073 ± 0.014	0.078 ± 0.018	0.527 ± 0.041
		MC-dropout	0.070 ± 0.012	0.075 ± 0.016	0.532 ± 0.037
HCP-Aging	rs-fMRI age group	CFIP	**0.073 ± 0.012**	**0.078 ± 0.015**	**0.553 ± 0.044**
		Gradient	0.058 ± 0.010	0.061 ± 0.013	0.512 ± 0.041
		SmoothGrad	0.070 ± 0.011	0.074 ± 0.014	0.522 ± 0.043
		Integrated Gradients	0.067 ± 0.011	0.071 ± 0.014	0.527 ± 0.041
		KernelSHAP	0.049 ± 0.013	0.055 ± 0.015	0.488 ± 0.046
		BootstrapSHAP	0.054 ± 0.012	0.058 ± 0.014	0.499 ± 0.048
		MC-dropout	0.056 ± 0.011	0.060 ± 0.013	0.508 ± 0.042

Table 12: Full multi-cohort faithfulness and stability results, reported as mean ± standard deviation over 5 folds. Table 4 summarizes the OASIS and HCP-Aging results.

Rather than evaluating sampled all-layer parameters in the original nonlinear network, we use the local tangent model

$$\tilde{f}_\alpha(z) = f_{\hat{\theta}}(z) + \psi_{\hat{\theta}}(z)^\top \alpha, \qquad \alpha = \theta - \hat{\theta}.$$

This is a finite-width local linearization around the trained model, related to neural tangent methods (Jacot et al., 2018; Lee et al., 2019), but it does not assume the infinite-width NTK regime. With $\Psi = [\psi_{\hat{\theta}}(x_1), \ldots, \psi_{\hat{\theta}}(x_N)]^\top$, $W = \mathrm{diag}(p_i(1-p_i))$, and $p_i = \sigma(f_{\hat{\theta}}(x_i))$, the tangent Laplace precision is

$$\Lambda_{\tan} = \lambda I + \Psi^\top W \Psi, \qquad \Sigma_{\tan} \approx \Lambda_{\tan}^{-1}, \tag{13}$$

using diagonal or structured approximations when needed.

Benchmark / Task	Method	VarCorr-S ↑	VarRelErr ↓	Ens-Cov@90 → .90	CovErr ↓	Sign-Brier ↓
ADNI CN vs. AD	CFIP	**0.910 ± 0.031**	**0.632 ± 0.156**	**0.887 ± 0.027**	**0.013 ± 0.027**	**0.044 ± 0.010**
	MC-dropout	0.537 ± 0.139	0.698 ± 0.112	0.812 ± 0.046	0.088 ± 0.046	0.079 ± 0.018
	BootstrapSHAP	0.179 ± 0.008	1.649 ± 0.197	0.741 ± 0.058	0.159 ± 0.058	0.124 ± 0.025
ADNI AD vs. MCI	CFIP	**0.897 ± 0.038**	**0.674 ± 0.161**	**0.876 ± 0.031**	**0.024 ± 0.031**	**0.052 ± 0.012**
	MC-dropout	0.419 ± 0.177	0.704 ± 0.175	0.793 ± 0.052	0.107 ± 0.052	0.088 ± 0.021
	BootstrapSHAP	0.166 ± 0.013	1.691 ± 0.173	0.724 ± 0.061	0.176 ± 0.061	0.132 ± 0.028
ADNI CN vs. MCI	CFIP	**0.888 ± 0.040**	**0.667 ± 0.171**	**0.872 ± 0.034**	**0.028 ± 0.034**	**0.058 ± 0.014**
	MC-dropout	0.401 ± 0.181	0.775 ± 0.115	0.781 ± 0.055	0.119 ± 0.055	0.094 ± 0.023
	BootstrapSHAP	0.161 ± 0.010	1.520 ± 0.063	0.708 ± 0.064	0.192 ± 0.064	0.139 ± 0.030
OASIS CN vs. dementia	CFIP	**0.842 ± 0.050**	**0.701 ± 0.138**	**0.868 ± 0.036**	**0.032 ± 0.036**	**0.061 ± 0.015**
	MC-dropout	0.498 ± 0.121	0.773 ± 0.129	0.805 ± 0.049	0.095 ± 0.049	0.091 ± 0.020
	BootstrapSHAP	0.211 ± 0.052	1.428 ± 0.164	0.733 ± 0.060	0.167 ± 0.060	0.129 ± 0.027
HCP-Aging rs-fMRI age group	CFIP	**0.756 ± 0.063**	**0.754 ± 0.151**	**0.851 ± 0.044**	**0.049 ± 0.044**	**0.073 ± 0.018**
	MC-dropout	0.438 ± 0.134	0.826 ± 0.141	0.789 ± 0.057	0.111 ± 0.057	0.105 ± 0.026
	BootstrapSHAP	0.252 ± 0.061	1.382 ± 0.149	0.719 ± 0.065	0.181 ± 0.065	0.145 ± 0.031

Table 13: Full uncertainty reliability and calibration results. VarCorr-S and VarRelErr compare each method's attribution-variance map with ensemble-retraining variability. Ens-Cov@90 measures coverage of retrained-model influence samples by 90% attribution intervals. CovErr is $|\mathrm{Cov@90} - 0.90|$. Sign-Brier compares predicted sign confidence with empirical sign frequency across retrained models. The ensemble provides the empirical retraining-variability reference.

Functional	Uncertainty	AOPC ↑	AOPIC ↑	Method
Ablation	Laplace	**0.177 ± 0.019**	**0.257 ± 0.021**	CFIP
Ablation	Ensemble	0.172 ± 0.020	0.255 ± 0.022	Ensemble-Ablation
Gradient	Laplace	0.127 ± 0.016	0.232 ± 0.019	CFIP-Grad
Gradient	Deterministic	0.126 ± 0.016	0.230 ± 0.018	Gradient

Table 14: Targeted ablation and insertion analysis on ADNI CN vs. AD. AOPIC is an insertion metric and therefore does not directly reward deletion-based ablation.

Prior translation and posterior center. The original prior is $\theta \sim \mathcal{N}(0, \lambda^{-1}I)$, so the perturbation prior is $\alpha \sim \mathcal{N}(-\hat{\theta}, \lambda^{-1}I)$. For Bernoulli cross-entropy, the tangent-model negative log posterior is

$$F_{\tan}(\alpha) = \sum_i \ell\big(y_i, f_{\hat{\theta}}(x_i) + \psi_{\hat{\theta}}(x_i)^\top \alpha\big) + \frac{\lambda}{2}\|\hat{\theta} + \alpha\|^2.$$

At a stationary solution of Eq. (3), with the same loss normalization and $\lambda > 0$,

$$\nabla F_{\tan}(0) = \Psi^\top(p - y) + \lambda\hat{\theta} = 0, \qquad \nabla^2 F_{\tan}(0) = \lambda I + \Psi^\top W \Psi.$$

The tangent objective is strictly convex, so its unique mode is zero. Its Laplace approximation therefore has a zero-centered perturbation posterior and the precision in Eq. (13). The Gaussian approximation is applied to this posterior; the translated prior retains mean $-\hat{\theta}$.

For ROI k, define the tangent influence feature

$$a_k(x) = \psi_{\hat{\theta}}(x) - \psi_{\hat{\theta}}(x^{(-k)}).$$

Stacking $a_k(x)^\top$ as rows of $A_{\tan}(x)$ gives an affine influence model,

$$\tilde{I}(x; \alpha) = I(x; \hat{\theta}) + A_{\tan}(x)\alpha.$$

Therefore, for $\alpha \sim \mathcal{N}(0, \Sigma_{\tan})$,

$$\tilde{I}(x) \mid \mathcal{D} \sim \mathcal{N}\Big(I(x; \hat{\theta}), A_{\tan}(x)\Sigma_{\tan}A_{\tan}(x)^\top\Big). \tag{14}$$

Thus full-network tangent CFIP uses all-layer gradient features while preserving closed-form inference in the tangent model. The nonlinear approximation is assessed empirically in Section 5.5; the moment guarantee in Theorem 3 requires the stated global remainder condition.

Variant	LapMC Spearman ↑	RelErr ↓	LapMC-Cov@90 → .90
ADNI LR exact	0.9998 ± 0.0000	0.026 ± 0.012	0.901 ± 0.006
ADNI MLP diag	0.9998 ± 0.0000	0.031 ± 0.014	0.899 ± 0.008
ADNI MLP KFAC	0.9997 ± 0.0000	0.047 ± 0.013	0.896 ± 0.010
OASIS MLP diag	0.9995 ± 0.0001	0.039 ± 0.015	0.895 ± 0.011
HCP-Aging sMRI MLP diag	0.9993 ± 0.0002	0.044 ± 0.016	0.892 ± 0.013
HCP-Aging rs-fMRI MLP diag	0.9987 ± 0.0004	0.058 ± 0.019	0.887 ± 0.016

Table 15: Full Laplace–MC pushforward validation. Spearman measures rank agreement between analytic and Monte Carlo influence variances over subject–ROI pairs. RelErr measures normalized variance-map error. LapMC-Cov@90 measures whether analytic CFIP 90% intervals cover Monte Carlo influence samples drawn from the same Laplace posterior.

Baseline	Spearman ↑	$\Delta P(I_k > 0)$
Global mean	1.000	0.004
Cohort mean	0.993	0.021

Table 16: Baseline sensitivity for ADNI CN vs. AD. Spearman reports rank correlation of global ROI influence relative to the zero-baseline result. $\Delta P(I_k > 0)$ is the mean absolute change in analytic sign confidence.

D Additional Validation

We provide additional evaluations of the posterior approximation, ranking faithfulness, synthetic feature-importance recovery, and computational cost. We also examine deeper feature extractors and multiclass prediction.

D.1 HMC Reference and Prior Sensitivity

Table 5 compares Laplace influence variances with an HMC reference. Variance-rank correlations round to 1.000 across the three tasks. The median Laplace/HMC variance ratios are below one, and coverage ranges from 0.831 to 0.871. Thus, Laplace preserves variance ordering closely while estimating smaller variances and under-covering the HMC influence samples.

The prior-precision rule is $\lambda = n_{\text{fit}} 10^{-4}$, related to the optimizer's weight decay. Increasing the precision by a factor of ten preserves variance ordering ($\rho \approx 0.999$) but yields median variance ratios of 0.340–0.389. Variance ranks are stable in this sweep, while the estimated variance scale changes substantially.

D.2 Q2: Ablations and Faithfulness Checks

In the paired CN–AD analysis, each subject's CFIP and IG ranking is evaluated through the same seed-0 predictor. Mean AOPC is 0.1752 for CFIP and 0.1661 for IG, with paired difference $+0.0091$, Wilcoxon $p = 6 \times 10^{-15}$, and paired-t $p = 7 \times 10^{-18}$. Both tests compare paired subject-level scores from this predictor.

Remove-and-retrain (ROAR) removes the top-k ranked features from both training and test inputs, then retrains the predictor. Lower AUC means that the removed features carried more predictive information. Maximum sensitivity measures attribution change under small input perturbations. Table 17 reports the five-fold CFIP–IG comparison alongside ROAR results for the other baselines.

Method	ROAR AUC $k = 10$	ROAR AUC $k = 20$	Max-sensitivity
CFIP	0.932 ± 0.012	0.920 ± 0.018	0.0087 ± 0.0006
Ensemble (10)	0.930 ± 0.012	0.917 ± 0.018	—
SWAG	0.932 ± 0.012	0.920 ± 0.018	—
Variational BNN	0.934 ± 0.013	0.922 ± 0.019	—
IG	0.935 ± 0.012	0.923 ± 0.020	0.0090 ± 0.0006

Table 17: Remove-and-retrain and sensitivity evaluation on CN–AD; lower is better. CFIP–IG results are mean $\pm$ SD over five folds. Dashes indicate N/A.

Table 3 reports the $k = 10$ ROAR results; Table 17 retains both removal levels. Insertion AOPIC and comparisons with CFIP-Grad and Ensemble-Ablation appear in Appendix B.6.

D.3 Bayesian Baseline Comparison

Table 3 gives the CN–AD comparison with a ten-model deep ensemble, SWAG, and a variational BNN. Table 18 reports the other two diagnosis tasks. Appendices B.4 and B.5 give the separate multi-cohort attribution benchmark.

D.4 Synthetic Feature-Importance Recovery

The Friedman #1 regression and INVASE Syn1–Syn6 classification studies use five seeds, a 64-ReLU feature extractor, a complete 65-dimensional head refit, and a five-network ensemble. The complete-head setting differs from the 64-weight diagonal timing path. Instance AUROC assesses feature-importance recovery; MSE and R^2 compare estimated feature effects with the generator's reference effects, measuring the accuracy of feature-importance recovery.

Table 1 in the main text reports the SHAP comparison and the INVASE results alongside CFIP and the ensemble.

Task	Method	AOPC ↑	Jacc@10 ↑	Spearman ↑	Var-ρ ↑	Cov@90
AD–MCI	CFIP	.066 ± .008	.527 ± .077	.694 ± .054	.908	.893
AD–MCI	Deep ensemble (10)	.059 ± .015	.517 ± .077	.694 ± .054	—	.911
AD–MCI	SWAG	.065 ± .008	.512 ± .076	.696 ± .055	.736	.816
AD–MCI	Variational BNN	.060 ± .009	.502 ± .009	.625 ± .016	.904	.860
CN–MCI	CFIP	.077 ± .021	.534 ± .039	.723 ± .034	.904	.708
CN–MCI	Deep ensemble (10)	.081 ± .023	.532 ± .039	.720 ± .034	—	.909
CN–MCI	SWAG	.079 ± .018	.554 ± .053	.729 ± .032	.735	.730
CN–MCI	Variational BNN	.074 ± .017	.062 ± .012	.114 ± .021	.893	.760

Table 18: Bayesian comparisons on AD–MCI and CN–MCI, completing the CN–AD comparison in Table 3. Cov@90 targets 0.90; Var-ρ and Cov@90 are point summaries. Dashes indicate N/A.

Training size	Method	Instance AUROC	Importance MSE	Importance R^2
500	CFIP	.922 ± .013	.469 ± .464	.931 ± .068
500	Ensemble	.919 ± .011	.466 ± .149	.930 ± .023
2000	CFIP	.940 ± .002	.136 ± .019	.980 ± .003
2000	Ensemble	.948 ± .002	.139 ± .008	.979 ± .001
10000	CFIP	.965 ± .002	.099 ± .013	.984 ± .002
10000	Ensemble	.969 ± .002	.095 ± .008	.986 ± .001

Table 19: Friedman feature-importance recovery over five seeds (mean ± SD). Global AUROC is reported as 1.000 at every size.

For INVASE, feature-importance ECE (lower is better) is 0.063 ± 0.002 for CFIP and 0.058 ± 0.002 for the ensemble. These calibration results are also reported in Section 5.2.

D.5 Depth and Frozen-Extractor Uncertainty

We examine how last-layer uncertainty changes with network depth, using an ensemble of fifteen fits with different seeds on one dataset. Table 2 compares CFIP and ensemble estimates with a synthetic variance reference. The reference variance grows as the extractor deepens, while last-layer CFIP varies less. This gap reflects the extractor uncertainty omitted when inference is restricted to the last layer.

D.6 Multiclass and Regression Extensions

For a multiclass predictor, define the influence on class c as $I_{kc}(x) = f_c(x) - f_c(x^{(-k)})$. With a frozen feature extractor it is affine in the class-c head weights, so the same Gaussian pushforward uses that block of the last-layer covariance. We apply this construction to class logits; differences in softmax probabilities require a nonlinear approximation. Table 9 reports the CN/MCI/AD evaluation.

For scalar regression with an affine output head, the output-difference functional is also linear in its weights. With a Gaussian likelihood and Gaussian prior, the head posterior is Gaussian conditional on the fixed extractor and noise variance, yielding an exact Gaussian influence law under that conditional model.

D.7 A Reliability Example

Across ten retrainings of CN–AD fold 0, fourteen ROIs appear in the top-ten lists: six always appear and eight appear intermittently. The intermittently selected group has 1.24 times the uncertainty of the always-selected group ($p = 8 \times 10^{-9}$); its sign-resolution fraction is 72% versus 92% ($p = 6 \times 10^{-5}$). The feature labeled ST52TA appears in one of the ten lists and has a sign-resolution fraction of 51%. ROIs selected less consistently have greater uncertainty and less resolved effect signs. The influence posterior therefore adds reliability information to the point importance ranking.

D.8 Explanation-Phase Runtime and Memory

CFIP uses the 64-weight diagonal path in this comparison. MC-dropout uses fifty gradient samples, BootstrapSHAP uses KernelSHAP, and the ensemble has ten models. Table 7 separates one-time explanation precomputation from per-subject explanation time. Timings cover the explanation phase and exclude model training; they use the implementation and sampling budgets described above.