COUNTERFACTUAL SUPPORT FIELDS FOR TASK-PRESERVING MEDICAL IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image reconstruction is commonly optimised for whole-image fidelity, although errors of similar magnitude can have different diagnostic consequences depending on their location. Under aggressive radial MR undersampling and sparse-view CT, lesion structures may remain degraded despite high global image quality. We propose **CSF-Recon**, a diagnosis-oriented framework that uses a *counterfactual support field (CSF)* to coordinate spatially targeted reconstruction and diagnostic evidence aggregation. Our analysis bounds diagnostic-margin degradation by reconstruction error weighted by spatial decision sensitivity, motivating preferential restoration of diagnostically relevant regions. Counterfactual alignment trains a reference teacher on fully sampled images to discourage reliance on decision-changing evidence outside annotated lesions. Its lesion-restricted class responses supervise a student field that gates microstructure restoration, local refinement, and local diagnostic pooling while retaining global context. The student predicts the field from a physics-driven coarse reconstruction, enabling spatial guidance directly from incomplete measurements. The teacher and lesion annotations are required only during training. On fastMRI+ and LIDC-IDRI at 16-fold acceleration, CSF-Recon achieves the highest mean accuracy and AUC among evaluated methods without maximising whole-image PSNR or SSIM. These results support spatially selective restoration as a practical approach to preserving diagnostic information under incomplete measurements.

1 INTRODUCTION

Undersampled magnetic resonance (MR) imaging and sparse-view computed tomography (CT) recover images from incomplete measurements. Learning-based methods improve reconstruction by combining acquisition physics with learned image priors (Schlemper et al., 2018; Hammernik et al., 2018; Sriram et al., 2020; Adler & Öktem, 2018). These methods are commonly optimised and evaluated using whole-image fidelity measures such as PSNR and SSIM. Such metrics quantify aggregate reconstruction error but do not explicitly distinguish its diagnostic consequences across image locations. Human-observer studies in MRI and AI-observer studies in sparse-view CT report corresponding discrepancies between image quality and task performance (Herman et al., 2024; Tivnan et al., 2025). This motivates evaluating reconstruction by how well it preserves diagnostically relevant structure, alongside conventional measures of image fidelity.

Task-adapted reconstruction addresses downstream requirements by jointly optimising image recovery and task prediction (Adler et al., 2022; Wu et al., 2024). However, an image-level task loss does not explicitly specify where additional restoration should be concentrated. Moreover, medical-image classifiers can exploit confounding features outside the pathology (DeGrave et al., 2021). Using their attention maps as reconstruction guides may therefore reinforce evidence that is predictive but anatomically unreliable (Arun et al., 2021). The challenge is to obtain spatial guidance that reflects diagnostic relevance and to use it consistently during reconstruction and prediction. Such guidance must remain available from incomplete measurements at inference. Detailed related work is provided in Appendix B.

We propose **CSF-Recon**, a diagnosis-oriented framework built around a *counterfactual support field (CSF)*. A reference teacher learns from fully sampled images through counterfactual alignment, which discourages decision-changing evidence outside the available lesion annotations

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.

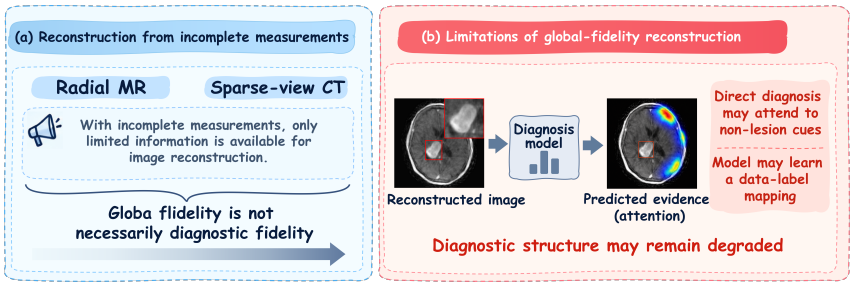

Figure 1: **Motivation for diagnosis-oriented reconstruction.** Retrospective radial sampling of MR k-space and simulated sparse-view CT projections provide incomplete measurements, limiting the information available for reconstruction. Under these conditions, reconstruction errors of similar magnitude can have different diagnostic consequences depending on where they occur. Whole-image fidelity objectives do not explicitly encode this spatial heterogeneity. CSF-Recon uses counterfactually aligned spatial evidence to guide restoration toward diagnostically relevant regions while retaining global image context.

(Liu et al., 2025). Its aligned class-response map is restricted to the annotated region and transferred to the student as spatial supervision. The field serves as an operational proxy for diagnostically relevant evidence, without assuming exact causal attribution. An independent frozen diagnostic probe evaluates the field's correspondence with spatial decision sensitivity (Appendix C.8). The teacher and lesion annotations are required only during training.

Given a physics-driven coarse reconstruction, the student predicts the support field and uses it to gate high-frequency microstructure restoration and iterative local refinement. The final reconstruction is re-encoded for diagnosis by combining support-weighted local features with global context. The same spatial variable therefore connects *where image detail is restored* with *where local diagnostic evidence is aggregated*. The downstream design is shared across radial MRI and sparse-view CT, with modality-specific forward models and coarse reconstructors. Figure 1 illustrates the motivation for prioritising diagnostically relevant structure under incomplete measurements.

Our analysis provides a decision-preservation bound in which reconstruction errors are weighted by spatial diagnostic sensitivity. A complementary local analysis relates counterfactual optimisation to diagnostic-margin-sensitive directions under explicit assumptions. These results motivate spatial allocation without equating the learned field with exact sensitivity. Experiments on fastMRI+ and LIDC-IDRI evaluate fidelity and diagnosis alongside controlled ablations. Under a common frozen probe, CSF-Recon retains the highest mean accuracy and AUC among evaluated reconstructions, supporting diagnostic information preservation beyond its specialised head (Appendix C.9).

The contributions of this work are threefold:

- **Decision-sensitive reconstruction principle.** We derive a diagnostic-margin bound that accounts for the spatial sensitivity of reconstruction errors, providing a rationale for non-uniform restoration.
- **Unified spatial guidance for reconstruction and diagnosis.** We introduce a counterfactual support field that couples microstructure restoration, local refinement, and diagnostic pooling through aligned teacher supervision.
- **Cross-modality evidence beyond pixel fidelity.** Experiments on radial MRI and sparse-view CT combine reconstruction and diagnostic evaluation with support-field controls, independent sensitivity analysis, and frozen diagnostic probes.

2 Method

2.1 Unified inverse-problem formulation

Let $d \in \{\mathrm{MR}, \mathrm{CT}\}$ index the modality, with real-valued reference image $\mathbf{x}_{\mathrm{gt}} \in \mathbb{R}^{H\times W}$, diagnostic label y, and lesion annotation $\mathbf{r}_{\mathrm{gt}} \in \{0,1\}^{H\times W}$. We use $\mathbf{u}_{\mathrm{CT}} = \mathbf{x}_{\mathrm{gt}}$, whereas $\mathbf{u}_{\mathrm{MR}} \in \mathbb{C}^{H\times W}$

denotes the complex MR object. Measurements follow

$$\mathbf{b}_d = \mathcal{A}_d \mathbf{u}_d + \boldsymbol{\epsilon}_d, \tag{1}$$

For C-coil MR, sensitivity encoding (Pruessmann et al., 1999) gives

$$\mathbf{b}_c = \mathcal{M}\mathcal{F}(\mathcal{S}_c \mathbf{u}_{\mathrm{MR}}) + \boldsymbol{\epsilon}_c, \qquad c = 1, \ldots, C, \tag{2}$$

where $\mathcal{S}_c$, $\mathcal{F}$ and $\mathcal{M}$ denote coil sensitivity, the centred Fourier transform and the radial sampling mask, respectively. Implementation details are provided in Appendix C. For CT, $\mathcal{A}_{\mathrm{CT}} = \mathcal{M}_v \mathcal{R}$ combines the Radon transform with sparse-view selection (Kak & Slaney, 2001). Let $\mathcal{I}_d$ denote the modality-specific initialisation: the multicoil adjoint for MR and backprojection or FBP for CT. The coarse reconstruction is

$$\mathbf{x}_c = R_\theta(\mathcal{I}_d(\mathbf{b}_d); \mathcal{A}_d). \tag{3}$$

Here, R_θ is a pretrained, frozen ISTA-Net-style unrolled network (Zhang & Ghanem, 2018); its optimisation interpretation and modality-specific architecture are detailed in Appendix C.2. The downstream student receives only the magnitude image $\mathbf{x}_c$, allowing a shared formulation across MR and CT. Figure 2 provides an overview of the framework, with the complete training and inference procedure detailed in Algorithm 1 of Appendix A.

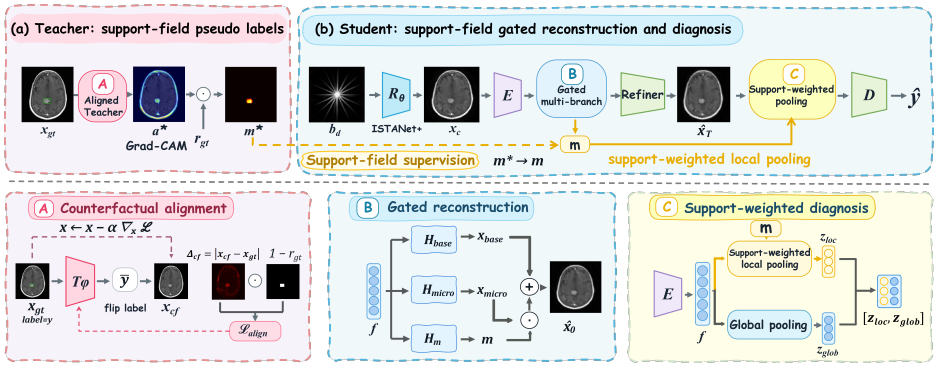

Figure 2: Overview of CSF-Recon. (a) The aligned teacher provides lesion-restricted Grad-CAM pseudo-labels $\mathbf{m}^\star$. (b) The student reconstructs measurements with ISTA-Net+, predicts support $\mathbf{m}$, and gates reconstruction and refinement before diagnosis from $\hat{\mathbf{x}}_T$. Insets show (A) counterfactual alignment, (B) base reconstruction plus support-gated microstructure, and (C) fusion of support-weighted local features with global pooled features. The teacher and lesion annotations are used only during training.

2.2 Counterfactual-alignment reference teacher

The reference teacher T_ϕ accesses fully sampled images only during training and predicts the diagnostic label from $\mathbf{x}_{\mathrm{gt}}$. Its purpose is not merely to obtain an accurate classifier, but to align the image evidence whose modification changes its decision with the lesion annotated by the radiologist.

We generate an image-space counterfactual using an established prediction-and-proximity objective (Wachter et al., 2018; Dhurandhar et al., 2018), as instantiated for medical decision alignment by Liu et al. (2025):

$$\mathbf{x}_{\mathrm{cf}}(\phi) = \arg\min_{\tilde{\mathbf{x}}} \ \mathrm{CE}(T_\phi(\tilde{\mathbf{x}}), \bar{y}) + \lambda_{\mathrm{cf}} d(\tilde{\mathbf{x}}, \mathbf{x}_{\mathrm{gt}}), \tag{4}$$

where $\bar{y} \neq y$ is the target counterfactual class and $d(\cdot, \cdot)$ discourages unnecessary deviation from the fully sampled reference. This inner objective follows standard counterfactual-explanation formulations and is not specific to CSF-Recon. Starting from $\mathbf{x}_{\mathrm{gt}}$, gradient descent iteratively updates $\tilde{\mathbf{x}}$ with the objective of flipping the predicted label. We define the counterfactual difference inline as $\boldsymbol{\Delta}_{\mathrm{cf}} = |\mathbf{x}_{\mathrm{cf}} - \mathbf{x}_{\mathrm{gt}}|$. Its lesion-exclusion alignment loss is

$$\mathcal{L}_{\mathrm{align}} = \frac{\|(\mathbf{1} - \mathbf{r}_{\mathrm{gt}}) \odot \boldsymbol{\Delta}_{\mathrm{cf}}\|_1}{\|\mathbf{1} - \mathbf{r}_{\mathrm{gt}}\|_1 + \epsilon}. \tag{5}$$

The difference identifies image regions whose modification can alter the diagnosis, while the alignment loss penalises such changes outside $\mathbf{r}_{\mathrm{gt}}$ and normalises them by the extra-lesional area.

The teacher is trained with

$$\mathcal{L}_{\mathrm{teacher}} = \mathrm{CE}(T_\phi(\mathbf{x}_{\mathrm{gt}}), y) + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}. \tag{6}$$

We differentiate the stationarity condition of the inner problem in Equation (4) implicitly (Pedregosa, 2016). A conjugate-gradient solve approximates the required inverse-Hessian-vector product without forming the inverse, yielding an approximate gradient for the outer objective in Equation (6). Minimising $\mathcal{L}_{\mathrm{align}}$ therefore discourages diagnosis-flipping evidence outside the available lesion annotation region and encourages the teacher's decision-sensitive evidence to align with the radiologist annotation. Appendix C.7 compares its CAM with that of an otherwise identical teacher trained using classification loss alone.

2.3 Support-field pseudo-labels

After counterfactual-alignment training, the teacher is frozen. We compute the true-class activation map $\mathbf{a}^\star$ from T_ϕ using Grad-CAM (Selvaraju et al., 2017) and rescale it to $[0, 1]$. The support-field pseudo-label is constructed by restricting this class-related response to the available lesion annotation region:

$$\mathbf{m}^\star = \mathbf{r}_{\mathrm{gt}} \odot \mathbf{a}^\star. \tag{7}$$

Here, $\mathbf{r}_{\mathrm{gt}}$ denotes the modality-specific lesion annotation available for supervision: in MR it is the binary region obtained by rasterising the union of radiologist-provided bounding boxes, whereas in CT it is the pixel-level region derived from the multi-reader lesion contours. Thus, $\mathbf{r}_{\mathrm{gt}}$ need not have the same annotation granularity across modalities, while $\mathbf{m}^\star$ consistently retains the class-related evidence within the available annotated region. The latter supervises the student support field $\mathbf{m}$.

2.4 Support-field gated structured reconstruction

The shared representation backbone E is a lightweight U-Net (Ronneberger et al., 2015) that preserves pixel alignment. This conservative choice controls backbone capacity, allowing improvements to be attributed to the proposed support-field mechanism rather than to a more advanced feature extractor. It first encodes the coarse reconstruction:

$$\mathbf{f}_c = E(\mathbf{x}_c). \tag{8}$$

Three lightweight heads then predict base anatomy, candidate microstructure and the student support field:

$$\mathbf{x}_{\mathrm{base}} = \mathcal{L}_\sigma(\mathbf{x}_c + H_{\mathrm{base}}(\mathbf{f}_c)), \quad \mathbf{x}_{\mathrm{micro}} = \mathcal{H}_\tau(H_{\mathrm{micro}}(\mathbf{f}_c)), \quad \mathbf{m} = \sigma(H_m(\mathbf{f}_c)). \tag{9}$$

Here, $\mathcal{L}_\sigma$ is a Gaussian low-pass operator and $\mathcal{H}_\tau = \mathcal{I} - \mathcal{L}_\tau$ is a high-pass operator. Our support-gated fusion defines the first-stage output as

$$\hat{\mathbf{x}}_0 = \mathbf{x}_{\mathrm{base}} + \mathbf{m} \odot \mathbf{x}_{\mathrm{micro}}. \tag{10}$$

Figure 3 schematically illustrates the intended spatial allocation of recovery: diagnostically relevant regions receive stronger local restoration than the background. The low-pass base branch preserves broad anatomy, while the support-field gate selectively injects candidate high-frequency structure within these regions. This design aims to retain lesion boundaries and microstructure without distributing high-frequency corrections indiscriminately across the image. The decomposition does not require $\mathbf{x}_{\mathrm{micro}}$ to be visually interpretable on its own; it represents the correction available for support-guided injection.

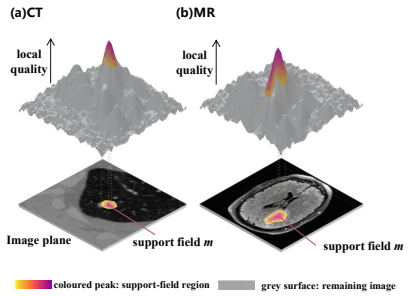

Figure 3: Schematic support-guided recovery in CT (a) and MR (b). Peaks denote support-prioritized restoration, not measured quality.

After this initial fusion, a lightweight iterative refiner applies support-gated residual corrections to $\hat{\mathbf{x}}_0$ and produces the final reconstruction $\hat{\mathbf{x}}_T$. The gate restricts refinement to diagnostically relevant regions; the iterative update, inputs and implementation are detailed in Appendix C.3.

2.5 SUPPORT-FIELD GATED DIAGNOSIS

Diagnosis is performed on the final reconstruction $\hat{\mathbf{x}}_T$, which is re-encoded by the shared encoder:

$$\mathbf{f}_r = E(\hat{\mathbf{x}}_T), \qquad \mathbf{h} = H_{\text{diag}}(\mathbf{f}_r). \tag{11}$$

The predicted support field $\mathbf{m}$ controls local evidence pooling, while global pooling preserves anatomical context:

$$\mathbf{z}_{\text{loc}} = \frac{\sum_{i,j} m_{ij} \mathbf{h}_{:,ij}}{\sum_{i,j} m_{ij} + \epsilon}, \qquad \mathbf{z}_{\text{glob}} = \frac{1}{HW} \sum_{i,j} \mathbf{h}_{:,ij}. \tag{12}$$

The final prediction is

$$\hat{y} = D([\mathbf{z}_{\text{loc}}, \mathbf{z}_{\text{glob}}]). \tag{13}$$

Diagnosis is therefore coupled to the final reconstructed image. Support-weighted local evidence is fused with complementary global context, reflecting the hypothesis that diagnosis uses both support-selected detail and whole-image anatomy without assuming that $\mathbf{m}$ alone determines the decision. The rationale and scope are discussed in Appendix C.10, while the component and support-source ablations in Section 3.3 and the diagnostic probes in Appendix C.9 evaluate its spatial allocation and diagnostic role.

2.6 TRAINING OBJECTIVES

The predicted student support field $\mathbf{m}$ is supervised by the pseudo-label $\mathbf{m}^\star$ using BCE and a soft Dice overlap loss (Milletari et al., 2016) over the whole image:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{BCE}}(\mathbf{m}, \mathbf{m}^\star) + \mathcal{L}_{\text{Dice}}(\mathbf{m}, \mathbf{m}^\star). \tag{14}$$

Soft Dice mitigates the imbalance caused by the support region being much smaller than the background. Base anatomy and microstructure use direct supervision:

$$\mathcal{L}_{\text{base}} = \|\mathbf{x}_{\text{base}} - \mathcal{L}_\sigma(\mathbf{x}_{\text{gt}})\|_2^2, \qquad \mathcal{L}_{\text{micro}} = \frac{\|\mathbf{m}^\star \odot [\mathbf{x}_{\text{micro}} - \mathcal{H}_\tau(\mathbf{x}_{\text{gt}})]\|_2^2}{\|\mathbf{m}^\star\|_1 + \epsilon}. \tag{15}$$

The final diagnostic loss is class-weighted cross-entropy, $\mathcal{L}_{\text{diag}} = \text{CE}(\hat{y}, y)$. The student and diagnostic branches optimise

$$\mathcal{L}_{\text{student}} = \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{base}}\mathcal{L}_{\text{base}} + \lambda_{\text{micro}}\mathcal{L}_{\text{micro}} + \lambda_{\text{diag}}\mathcal{L}_{\text{diag}}. \tag{16}$$

The refiner is optimised separately using the support-field-weighted mean-squared error $\mathcal{L}_{\text{ref}}$, whose exact form is provided in Appendix C.3. The pretrained coarse reconstructor R_θ remains frozen. The first-stage reconstruction, diagnosis and refiner objectives are optimised with their corresponding parameter groups so that each module is trained for its stated role.

2.7 DECISION-SENSITIVE RECONSTRUCTION

Standard unweighted pixel-wise reconstruction objectives treat spatial errors uniformly, although errors at different locations can affect a downstream decision differently. For a fixed binary diagnostic network with logits $\mathcal{F}_y(\mathbf{x})$ and $\mathcal{F}_{\bar{y}}(\mathbf{x})$, define the diagnostic margin $\mathcal{G}(\mathbf{x}) = \mathcal{F}_y(\mathbf{x}) - \mathcal{F}_{\bar{y}}(\mathbf{x})$. Let $\mathbf{x}^\star$ be the fully sampled reference, $\hat{\mathbf{x}} = \mathbf{x}^\star + \mathbf{e}$ a reconstruction, and assume $\mathcal{G}(\mathbf{x}^\star) > 0$. Define the path-averaged decision sensitivity $\bar{s}_i(\mathbf{e}) = \int_0^1 |\partial_i \mathcal{G}(\mathbf{x}^\star + t\mathbf{e})|\, dt$. Under the mild pathwise regularity conditions given in Appendix D.1,

$$\mathcal{G}(\hat{\mathbf{x}}) \geq \mathcal{G}(\mathbf{x}^\star) - \sum_i \bar{s}_i(\mathbf{e})|e_i|. \tag{17}$$

Hence, the reference decision is preserved whenever

$$\sum_i \bar{s}_i(\mathbf{e})|e_i| < \mathcal{G}(\mathbf{x}^\star). \tag{18}$$

Equation (17) shows that global reconstruction error does not explicitly encode spatially varying decision sensitivity, although global fidelity remains relevant. For sufficiently local perturbations, when the diagnostic gradient varies little along the reconstruction path, $\bar{s}_i(\mathbf{e}) \approx s_i = |\partial_i \mathcal{G}(\mathbf{x}^\star)|$.

This local view also explains the counterfactual objective used to align the reference teacher. For an unconstrained, locally linearized surrogate with standard unweighted binary softmax cross-entropy and isotropic quadratic ℓ_2 proximity,

$$\mathcal{L}_{\mathrm{sur}}(\boldsymbol{\delta}) = \log\big[1 + \exp\big(\mathcal{G}_0 + \mathbf{g}^\top \boldsymbol{\delta}\big)\big] + \frac{\lambda_{\mathrm{cf}}}{2}\|\boldsymbol{\delta}\|_2^2, \qquad \mathcal{G}_0 = \mathcal{G}(\mathbf{x}^\star), \quad \mathbf{g} = \nabla\mathcal{G}(\mathbf{x}^\star), \tag{19}$$

has a unique minimizer satisfying

$$\boldsymbol{\delta}^\star_{\mathrm{sur}} = -c\,\nabla\mathcal{G}(\mathbf{x}^\star), \qquad c > 0. \tag{20}$$

Thus, under this local surrogate, the regularized counterfactual objective follows diagnostic-margin-sensitive directions. Importantly, Equation (20) characterizes the local surrogate minimizer rather than an arbitrary finite-step or early-stopped counterfactual. Appendix D.3 gives the condition under which this minimizer reaches or crosses the locally linearized decision boundary.

This analysis provides a local rationale, not a causal-attribution guarantee: counterfactual alignment discourages extra-lesional decision-sensitive evidence, but the actual perturbation need not equal the surrogate minimizer. The student uses neither $|\nabla\mathcal{G}|$ nor counterfactual magnitude directly; it is guided by $\mathbf{m}^\star = \mathbf{r}_{\mathrm{gt}} \odot \mathbf{a}^\star$. The adequacy of this proxy is evaluated in Section 3.3 using shifted-mask, counterfactual-difference, GT, Plain Grad-CAM and Aligned Grad-CAM controls.

Finally, the local fixed-sensitivity quantity

$$\mathcal{B}_{\mathrm{local}}(\mathbf{e}) = \sum_i s_i |e_i|, \qquad s_i = |\partial_i \mathcal{G}(\mathbf{x}^\star)| \tag{21}$$

provides a design motivation for spatially non-uniform reconstruction. If only the absolute error at location i is reduced by $0 \le \Delta_i \le |e_i|$, while the sensitivities and other coordinate errors are fixed, then $\mathcal{B}_{\mathrm{local}}(\mathbf{e}) - \mathcal{B}_{\mathrm{local}}(\mathbf{e}') = s_i\Delta_i$. Thus, correcting regions with larger decision sensitivity yields a larger reduction in decision-sensitive error for comparable reducible errors. This motivates support-gated microstructure injection and residual refinement without claiming global optimality of the learned support field.

3 Experiments

3.1 Experimental setup

Datasets and preprocessing. For the MR task, we construct a binary subset of fastMRI+ (Zbontar et al., 2018; Zhao et al., 2022) by treating *Nonspecific lesion* as class 0 and merging *Mass* and *Extra-axial mass* as class 1. Volumes containing both classes are excluded, resulting in 168 volumes and 870 annotated slices. Multicoil images are resized to 320×320 and normalised to $[0, 1]$; lesion bounding boxes are converted into ROI masks. For CT, LIDC-IDRI images (Armato III et al., 2011) are normalised using a lung window centred at -600 HU with a width of 1500 HU. Nodules smaller than $7\,\mathrm{mm}$ are removed, and 96×96 patches are extracted. The median malignancy rating determines the binary label: scores ≤ 3 are benign and scores > 3 are malignant. The processed dataset is class balanced and split at the patient level into 7,653 training, 1,504 validation and 1,537 test samples. Further preprocessing and reconstruction details are provided in Appendix C.

Experimental settings. MR data are retrospectively undersampled using golden-angle radial masks, whereas CT measurements are generated from 720-view parallel-beam projections. Both MR and CT experiments use an acceleration factor of 16; for CT, this corresponds to retaining 45 of the 720 projection views. In the teacher network, the implicit counterfactual variable is optimised for 150 steps with a step size of 0.06. For both modalities, the low-pass operator uses a 13×13 Gaussian kernel with $\sigma = 2.0$, while the high-pass operator uses a 9×9 kernel with $\sigma = 1.5$. The local refiner performs six iterative correction steps. Diagnostic gradients are detached from the shared encoder and support predictor. The main quantitative results are evaluated at the slice level, using ROI-PSNR and ROI-SSIM for lesion-region reconstruction fidelity and accuracy and AUC for diagnostic performance. All experiments are independently trained with three random seeds, and results are reported as mean $\pm$ standard deviation.

3.2 BASELINE COMPARISONS

We compare zero-filled/FBP reconstruction, U-Net post-processing (Ronneberger et al., 2015), learned primal-dual reconstruction, PnP-HQS (Venkatakrishnan et al., 2013), MoDL (Aggarwal et al., 2019), ISTA-Net+, and a task-oriented U-Net joint baseline. U-Net joint optimises $\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{task}} \mathcal{L}_{\text{diag}}$ without support-field supervision or spatial gating; all methods use matched inputs and modality-specific training. For each baseline, the same classifier is trained on its reconstructions to report accuracy and AUC.

Table 1: Baseline comparison on fastMRI+ and LIDC-IDRI at an acceleration factor of 16. Results are mean ± standard deviation over three seeds. Bold values are best and underlined values are second best; CSF-Recon achieves the best accuracy and AUC in both modalities.

	MR fastMRI+			CT LIDC		
Method	ROI-PSNR	Acc.	AUC	ROI-PSNR	Acc.	AUC
Zero-filled / FBP	22.79 ± 0.00	0.593 ± 0.043	0.587 ± 0.153	19.52 ± 0.00	0.631 ± 0.040	0.617 ± 0.096
U-Net	23.33 ± 0.20	0.615 ± 0.048	0.759 ± 0.033	28.09 ± 0.32	0.639 ± 0.009	0.669 ± 0.026
U-Net joint	20.37 ± 0.52	0.752 ± 0.045	0.796 ± 0.036	24.26 ± 2.07	0.671 ± 0.011	0.744 ± 0.014
Learned primal-dual (LPD)	26.22 ± 0.01	0.667 ± 0.030	0.669 ± 0.118	**33.75 ± 0.25**	0.620 ± 0.027	0.660 ± 0.006
PnP-HQS	**27.36 ± 0.00**	0.558 ± 0.192	0.744 ± 0.123	21.33 ± 0.00	0.619 ± 0.025	0.665 ± 0.011
MoDL	27.02 ± 0.07	0.689 ± 0.030	0.697 ± 0.174	30.25 ± 0.06	0.605 ± 0.035	0.658 ± 0.018
ISTA-Net+	27.13 ± 0.00	0.673 ± 0.042	0.709 ± 0.143	30.52 ± 0.00	0.625 ± 0.030	0.674 ± 0.017
CSF-Recon	27.20 ± 0.17	**0.763 ± 0.010**	**0.825 ± 0.089**	30.63 ± 0.34	**0.734 ± 0.010**	**0.821 ± 0.018**

Table 1 shows that PnP-HQS and LPD achieve the highest MR and CT ROI-PSNR, respectively, whereas CSF-Recon attains the best accuracy and AUC in both modalities while ranking second in ROI-PSNR (Figure 4). Joint U-Net training also improves diagnostic performance despite lower ROI-PSNR. Figure 5 suggests that CSF-Recon better preserves lesion boundaries and intralesional texture. Although PnP-HQS attains the highest MR ROI-PSNR, its reconstructions are visibly oversmoothed, an effect that is even more pronounced in CT. Appendix C.5 reports case-level accuracy and paired bootstrap analysis, complementing the slice-level results in Table 1. Additional PSNR and SSIM results are reported in Tables 4 and 5 of Appendix C.

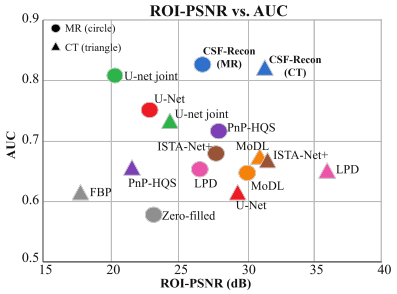

Figure 4: ROI fidelity versus diagnostic utility. Circles denote MR and triangles CT; higher ROI-PSNR does not necessarily imply higher AUC.

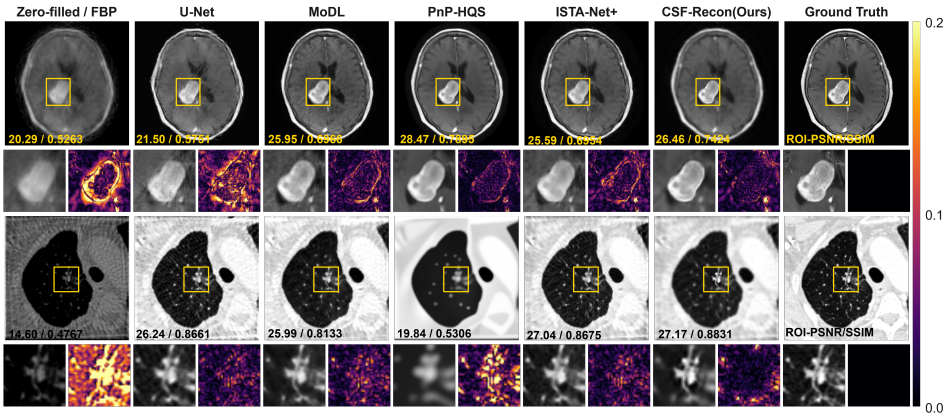

Figure 5: Qualitative MR and CT reconstruction at an acceleration factor of 16. CSF-Recon uses support-gated high-frequency restoration and local refinement to suppress streak-like artefacts while preserving lesion boundaries and internal contrast.

With a common frozen diagnostic probe, CSF-Recon also achieves the highest mean accuracy and AUC among evaluated reconstruction methods (Appendix C.9, Table 10).

3.3 Ablation Studies

Main ablation. Table 2 evaluates support-field gated reconstruction, local refinement and support-field gated diagnosis. The support field enables spatially selective recovery, and the refiner improves ROI-PSNR and ROI-SSIM. The diagnostic gate mainly raises accuracy and AUC: despite slightly lower local fidelity, the third configuration ranks second on both diagnostic metrics. The full model performs best overall; corresponding visualisations are provided in Appendix C (Figure 9).

Table 2: Progressive component ablation of CSF-Recon on MR and CT.

			MR fastMRI+				CT LIDC			
Support field	Refiner	Diag. gate	ROI-PSNR	ROI-SSIM	Acc.	AUC	ROI-PSNR	ROI-SSIM	Acc.	AUC
×	×	×	26.79 ± 0.00	0.9003 ± 0.0000	0.679 ± 0.073	0.764 ± 0.065	30.28 ± 0.00	0.9735 ± 0.0000	0.650 ± 0.015	0.682 ± 0.005
✓	×	×	26.19 ± 0.10	0.8909 ± 0.0027	0.728 ± 0.065	0.731 ± 0.035	28.53 ± 0.05	0.9636 ± 0.0006	0.712 ± 0.025	0.810 ± 0.020
✓	×	✓	26.16 ± 0.06	0.8902 ± 0.0009	0.731 ± 0.040	0.791 ± 0.006	28.77 ± 0.10	0.9632 ± 0.0017	0.729 ± 0.013	0.819 ± 0.018
✓	✓	×	27.17 ± 0.01	0.9101 ± 0.0033	0.718 ± 0.029	0.783 ± 0.006	30.55 ± 0.14	0.9752 ± 0.0006	0.717 ± 0.016	0.803 ± 0.008
✓	✓	✓	$\mathbf{27.20 \pm 0.17}$	$\mathbf{0.9112 \pm 0.0041}$	$\mathbf{0.763 \pm 0.010}$	$\mathbf{0.825 \pm 0.089}$	$\mathbf{30.63 \pm 0.34}$	$\mathbf{0.9753 \pm 0.0005}$	$\mathbf{0.734 \pm 0.010}$	$\mathbf{0.821 \pm 0.018}$

Support-field source ablation. We compare five support-field sources while keeping the student reconstruction and diagnostic pipelines fixed: a spatially random lesion mask, the annotation-restricted counterfactual difference, the lesion annotation, Plain Grad-CAM, and Aligned Grad-CAM (Ours). Plain Grad-CAM is produced by an otherwise identical teacher trained with classification loss only, whereas the aligned teacher additionally uses $\mathcal{L}_{\text{align}}$. This isolates whether counterfactual alignment makes classifier attention more suitable for reconstruction guidance. Figure 6 provides representative MR and CT visualisations, and Table 3 reports the quantitative CT results.

Table 3: Quantitative CT support-field source ablation on LIDC-IDRI.

Source	ROI-PSNR ↑	ROI-SSIM ↑	Acc. ↑	AUC ↑
Random mask	29.47 ± 0.17	0.9704 ± 0.0014	0.542 ± 0.000	0.712 ± 0.034
$\Delta_{\text{cf}} \times \text{GT}$	30.50 ± 0.36	$\mathbf{0.9762 \pm 0.0006}$	0.601 ± 0.094	0.805 ± 0.021
GT	30.39 ± 0.37	0.9747 ± 0.0004	0.727 ± 0.011	0.813 ± 0.019
Plain Grad-CAM	30.07 ± 0.32	0.9742 ± 0.0006	0.658 ± 0.107	0.779 ± 0.031
Aligned Grad-CAM (Ours)	$\mathbf{30.63 \pm 0.34}$	0.9753 ± 0.0005	$\mathbf{0.734 \pm 0.010}$	$\mathbf{0.821 \pm 0.018}$

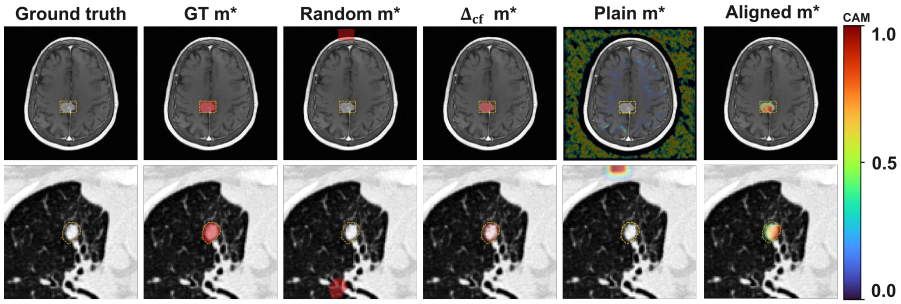

Figure 6: Qualitative support-field source ablation on MR and CT cases. Columns show the reference, GT, random mask, counterfactual difference, Plain CAM and Aligned CAM. The Plain $m^\star$ column displays the unmasked teacher CAM to reveal extra-lesional responses; both CAM sources use identical lesion restriction during training. Yellow dashed outlines indicate the lesion annotations in both modalities.

The shifted mask performs worst, confirming that correct lesion location matters, while GT weights the lesion uniformly and the counterfactual difference is spatially fragmented. Figure 6 shows weak

lesion responses in the plain teacher CAM and more coherent lesion-focused guidance after alignment. It also raises CT accuracy from 0.658 to 0.734 and AUC from 0.779 to 0.821, with concurrent gains in ROI fidelity. Together with the teacher-CAM comparison in Appendix C.7, these results indicate that counterfactual alignment is beneficial for constructing an effective reconstruction support field.

4 Discussion and Limitations

These results support spatially selective restoration as a practical strategy for preserving diagnostic information under severe undersampling. The decision-preservation bound links potential diagnostic-margin degradation to reconstruction errors weighted by spatial sensitivity, motivating preferential correction where errors matter most. Consistent with this view, CSF-Recon achieves the highest mean accuracy and AUC in both modalities at 16-fold acceleration without maximising ROI-PSNR (Table 1). This finding motivates evaluating reconstruction through both image fidelity and downstream diagnostic utility.

The support field gives this principle an operational form by coupling spatially targeted reconstruction with local diagnostic pooling. Component ablations support complementary roles for local refinement and diagnostic gating. Source ablations further indicate that lesion location and the quality of class-conditioned guidance both influence performance. Decision-sensitivity maps from an independent frozen probe agree more closely with aligned teacher maps and predicted student fields than with plain Grad-CAM (Appendix C.8). These observations support the field as an empirical spatial proxy, without establishing equivalence to input-gradient sensitivity or a causal interpretation.

The probe experiments help distinguish information retained in the reconstructed images from gains due to the diagnostic pathway. With a common classifier trained on reference images and frozen within each modality, CSF-Recon retains the highest mean accuracy and AUC among evaluated reconstructions (Appendix C.9). This supports a contribution from the reconstructed images themselves beyond the specialised diagnostic head. Matched probes additionally favour combining support-weighted local evidence with global features, indicating that targeted local information and anatomical context remain complementary.

The evidence remains limited to two binary tasks, retrospective radial MR undersampling, simulated sparse-view CT, and the evaluated acceleration setting. Training depends on lesion annotations and teacher response quality, although neither annotations nor the teacher are required at inference. Although the final refinement stage lacks an explicit data-consistency projection, using only a few refinement steps may help limit measurement inconsistency. Prospective multicentre studies, human-reader evaluation, and testing across acquisition conditions and diagnostic tasks are needed to assess generalisability. Reducing annotation dependence and incorporating post-refinement data consistency are further directions.

5 Conclusion

We introduced CSF-Recon, which uses a counterfactual support field to coordinate microstructure restoration, local refinement, and diagnostic evidence aggregation. Counterfactually aligned teacher responses provide training supervision, while the student predicts the field from a coarse reconstruction at inference. Decision-sensitive analysis motivates this spatial allocation, and experiments on radial MRI and sparse-view CT support its diagnostic utility. Ablations, independent sensitivity comparisons, and frozen diagnostic probes provide complementary evidence for the learned field and reconstructed images. Within the evaluated settings, these findings support spatial diagnostic relevance as a useful design principle for task-preserving reconstruction.

AI USE STATEMENT

Generative AI tools were used in this study to assist with translation, literature search and review, language editing, and programming. The authors reviewed all AI-assisted text and code, and verified the reported results and references against the original literature, underlying data, or other source materials. The authors take full responsibility for the final manuscript.

ETHICS STATEMENT

This study adheres to the ICLR Code of Ethics. It did not involve human participants or animal experiments. All medical-imaging datasets used in our experiments are publicly available under their respective research licences and were used in accordance with the applicable usage guidelines. The datasets contain no personally identifiable information, thereby preserving privacy and ethical compliance.

REPRODUCIBILITY STATEMENT

We have taken several measures to support the reproducibility of this study. Part of the experimental setup is described in the main text and appendices. The source code, dataset-preprocessing pipelines, model architectures, training hyperparameters and detailed documentation of the complete workflow will be packaged for anonymous upload and subsequently released together in an anonymous public code repository.

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. URL https://papers.nips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018. doi: 10.1109/TMI.2018.2799231.

Jonas Adler, Sebastian Lunz, Olivier Verdier, Carola-Bibiane Schönlieb, and Ozan Öktem. Task adapted reconstruction for inverse problems. *Inverse Problems*, 38(7):075006, 2022. doi: 10.1088/1361-6420/ac28ec.

Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 2019.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020. doi: 10.1073/pnas.1907377117.

Samuel G. Armato III et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 2011.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D. Li, and Jayashree Kalpathy-Cramer. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021. doi: 10.1148/ryai.2021200267.

Sayantan Bhadra, Varun A. Kelkar, Frank J. Brooks, and Mark A. Anastasio. On hallucinations in tomographic image reconstruction. *IEEE Transactions on Medical Imaging*, 40(11):3249–3260, 2021. doi: 10.1109/TMI.2021.3077857.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html.

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated MRI. *Medical Image Analysis*, 80:102479, 2022. doi: 10.1016/j.media.2022.102479.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619, 2021. doi: 10.1038/s42256-021-00338-7.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/c5ff2543b53f4cc0ad3819a36752467b-Abstract.html.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018. doi: 10.1002/mrm.26977.

Yuzhu He, Chunfeng Lian, Ruyi Xiao, Fangmao Ju, Chao Zou, Zongben Xu, and Jianhua Ma. Anatomy-aware deep unrolling for task-oriented acceleration of multi-contrast MRI. *IEEE Transactions on Medical Imaging*, 44(9):3832–3844, 2025. doi: 10.1109/TMI.2025.3568157.

Joshua D. Herman, Rachel E. Roca, Alexandra G. O'Neill, Marcus L. Wong, Sajan Goud Lingala, and Angel R. Pineda. Task-based assessment for neural networks: Evaluating undersampled MRI reconstructions based on human observer signal detection. *Journal of Medical Imaging*, 11(4):045503, 2024. doi: 10.1117/1.JMI.11.4.045503.

Kyong Hwan Jin et al. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 2017.

Avinash C. Kak and Malcolm Slaney. *Principles of Computerized Tomographic Imaging*. Society for Industrial and Applied Mathematics, 2001. doi: 10.1137/1.9780898719277.

Mingzhou Liu, Ching-Wen Lee, Xinwei Sun, Xueqing Yu, Yu Qiao, and Yizhou Wang. Learning causal alignment for reliable disease diagnosis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ozZG5FXuTV.

Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 2007.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 737–746, 2016. URL https://proceedings.mlr.press/v48/pedregosa16.html.

Klaas P. Pruessmann, Markus Weiger, Markus B. Scheidegger, and Peter Boesiger. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 42(5):952–962, 1999. URL https://pubmed.ncbi.nlm.nih.gov/10542355/.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8116–8126, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371.

Jo Schlemper, Jose Caballero, Joseph V. Hajnal, Anthony N. Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503, 2018. doi: 10.1109/TMI.2017.2760978.

Ramprasaath R. Selvaraju et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly: A counterfactual approach. *Medical Image Analysis*, 84: 102721, 2023. doi: 10.1016/j.media.2022.102721.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated MRI reconstruction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12262 of *Lecture Notes in Computer Science*, pp. 64–73. Springer, 2020. doi: 10.1007/978-3-030-59713-9_7.

Susu Sun, Stefano Woerner, Andreas Maier, Lisa M. Koch, and Christian F. Baumgartner. Inherently interpretable multi-label classification using class-specific counterfactuals. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 937–956, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL `https://proceedings.mlr.press/v70/sundararajan17a.html`.

Matthew Tivnan, Tzu-Cheng Lee, Ruoqiao Zhang, Kirsten Boedeker, Liang Cai, Jeremias Sulam, and J. Webster Stayman. Task-driven CT image quality optimization for low-contrast lesion detectability with tunable neural networks. In *Medical Imaging 2023: Physics of Medical Imaging*, volume 12463 of *Proceedings of SPIE*, pp. 124631M, 2023. doi: 10.1117/12.2653936.

Matthew Tivnan, Irene Désirée Kikkert, Dufan Wu, Kai Yang, Jelmer M. Wolterink, Quanzheng Li, and Rajiv Gupta. Task based evaluation of sparse view CT reconstruction techniques for intracranial hemorrhage diagnosis using an AI observer model. *Scientific Reports*, 15:26002, 2025. doi: 10.1038/s41598-025-11089-5.

Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *IEEE Global Conference on Signal and Information Processing*, 2013.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2018.

Zihui Wu, Tianwei Yin, Yu Sun, Robert Frost, Andre van der Kouwe, Adrian V. Dalca, and Katherine L. Bouman. Learning task-specific strategies for accelerated MRI. *IEEE Transactions on Computational Imaging*, 10:1040–1054, 2024. doi: 10.1109/TCI.2024.3410521.

Jure Zbontar et al. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018.

Jian Zhang and Bernard Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3929–3938, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhang_Learning_Deep_CNN_CVPR_2017_paper.html.

Ruiyang Zhao et al. fastMRI+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Scientific Data*, 2022.

A Training and Inference Workflow

Algorithm 1 summarises the two-stage training procedure and the student inference path. The acquisition object is $\mathbf{u}_{\mathrm{CT}} = \mathbf{x}_{\mathrm{gt}}$ for CT and a complex image $\mathbf{u}_{\mathrm{MR}}$ for multicoil MR; the latter is distinct from the magnitude reference $\mathbf{x}_{\mathrm{gt}}$. For CT, $\mathcal{A}_{\mathrm{CT}} = \mathcal{M}_v\mathcal{R}$ gives simulated sparse-view projections and $\mathcal{I}_{\mathrm{CT}}$ uses backprojection or FBP. For MR, $\mathcal{A}_{\mathrm{MR}}\mathbf{u} = \{\mathcal{M}\mathcal{F}(\mathcal{S}_c\mathbf{u})\}_{c=1}^{C}$ retrospectively masks measured multicoil k-space and $\mathcal{I}_{\mathrm{MR}} = \mathcal{A}_{\mathrm{MR}}^{\mathrm{H}}$. Fully sampled references, annotations and the teacher are used only during training.

Algorithm 1 CSF-Recon training and inference

Require: Training examples $(\mathbf{x}_{\mathrm{gt}}, \mathbf{r}_{\mathrm{gt}}, y, \mathbf{u}_d)$; modality d; $\mathcal{A}_d, \mathcal{I}_d$; pretrained frozen R_θ; refiner steps T.
Ensure: Trained student; test outputs $(\hat{\mathbf{x}}_T, \hat{y})$ from measurements $\mathbf{b}_d$.

Teacher stage (training only)
for each teacher-training batch **do**
  Compute $\mathbf{x}_{\mathrm{cf}}$ by Equation (4) with target $\bar{y} \neq y$.
  $\boldsymbol{\Delta}_{\mathrm{cf}} \leftarrow |\mathbf{x}_{\mathrm{cf}} - \mathbf{x}_{\mathrm{gt}}|$.
  Compute $\mathcal{L}_{\mathrm{align}}$ from $\boldsymbol{\Delta}_{\mathrm{cf}}$ and $\mathbf{r}_{\mathrm{gt}}$ via Equation (5).
  Update ϕ using $\mathcal{L}_{\mathrm{teacher}}$ in Equation (6).
end for
Freeze T_ϕ.
for each training reference **do**
  $\mathbf{a}^\star \leftarrow \mathrm{Normalize}(\mathrm{GradCAM}(T_\phi, \mathbf{x}_{\mathrm{gt}}, y))$.
  Cache $\mathbf{m}^\star \leftarrow \mathbf{r}_{\mathrm{gt}} \odot \mathbf{a}^\star$.
end for
Student stage (training; same forward path at inference)
for each student-training batch **do**
  $\mathbf{b}_d \leftarrow \mathcal{A}_d\mathbf{u}_d + \boldsymbol{\epsilon}_d$; $\mathbf{x}_{\mathrm{init}} \leftarrow \mathcal{I}_d(\mathbf{b}_d)$.
  $\mathbf{x}_c \leftarrow R_\theta(\mathbf{x}_{\mathrm{init}}; \mathcal{A}_d)$ (frozen R_θ).
  $\mathbf{f}_c \leftarrow E(\mathbf{x}_c)$.
  $\mathbf{x}_{\mathrm{base}} \leftarrow \mathcal{L}_\sigma(\mathbf{x}_c + H_{\mathrm{base}}(\mathbf{f}_c))$.
  $\mathbf{x}_{\mathrm{micro}} \leftarrow \mathcal{H}_\tau(H_{\mathrm{micro}}(\mathbf{f}_c))$.
  $\mathbf{m} \leftarrow \sigma(H_m(\mathbf{f}_c))$.
  $\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_{\mathrm{base}} + \mathbf{m} \odot \mathbf{x}_{\mathrm{micro}}$; $\mathbf{x}^{(0)} \leftarrow \hat{\mathbf{x}}_0$.
  for $t = 0, \ldots, T-1$ **do**
    Compute $\boldsymbol{\delta}_t$ from Equation (23).
    $\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} + \mathbf{m} \odot \boldsymbol{\delta}_t$.
  end for
  $\hat{\mathbf{x}}_T \leftarrow \mathbf{x}^{(T)}$; $\mathbf{h} \leftarrow H_{\mathrm{diag}}(E(\hat{\mathbf{x}}_T))$.
  $\mathbf{z}_{\mathrm{loc}} \leftarrow \mathrm{WeightedPool}(\mathbf{h}, \mathbf{m})$; $\mathbf{z}_{\mathrm{glob}} \leftarrow \mathrm{GlobalPool}(\mathbf{h})$.
  $\hat{y} \leftarrow D([\mathbf{z}_{\mathrm{loc}}, \mathbf{z}_{\mathrm{glob}}])$.
  Update student and diagnosis groups with $\mathcal{L}_{\mathrm{student}}$; refiner with $\mathcal{L}_{\mathrm{ref}}$.
end for
Inference: run the student forward path from $\mathbf{b}_d$ only; do not use T_ϕ, $\mathbf{x}_{\mathrm{gt}}$ or $\mathbf{r}_{\mathrm{gt}}$.

B Background and Related Work

B.1 Physics-driven reconstruction and learned image priors

Physics-driven reconstruction combines an acquisition model with image priors. Compressed sensing MRI uses sparsity regularisation and data fidelity (Lustig et al., 2007); cascaded CNNs alternate learned processing with data consistency (Schlemper et al., 2018). Variational networks, MoDL and

end-to-end VarNet learn regularisation within iterative reconstruction pipelines (Hammernik et al., 2018; Aggarwal et al., 2019; Sriram et al., 2020). For CT, FBPConvNet removes artefacts after analytic inversion (Jin et al., 2017), whereas learned primal-dual reconstruction learns measurement- and image-domain updates (Adler & Öktem, 2018). ISTA-Net learns transforms, thresholds and step sizes, with ISTA-Net+ operating on residuals (Zhang & Ghanem, 2018). Plug-and-play reconstruction inserts denoisers into iterative solvers (Venkatakrishnan et al., 2013), including CNN priors within half-quadratic splitting (Zhang et al., 2017). Score-based diffusion models instead combine generative image priors with measurement conditioning (Chung & Ye, 2022).

These approaches improve recovery from incomplete measurements, but image plausibility alone does not establish preservation of diagnostic evidence. Antun et al. document instabilities and missed small structures in tested reconstruction networks (Antun et al., 2020); Bhadra et al. analyse hallucinations through measurement-supported and prior-dependent image components (Bhadra et al., 2021). Both motivate evaluating local evidence alongside aggregate error. CSF-Recon adds spatially targeted restoration to a physics-based coarse reconstruction, with the support field complementing rather than replacing the acquisition model.

B.2 Task-oriented reconstruction and diagnostic evaluation

Task-oriented methods optimise recovery for downstream use. Adler et al. jointly train reconstruction and task estimators for classification and segmentation (Adler et al., 2022). TACKLE jointly learns MRI subsampling, reconstruction and prediction through pretraining and task-specific fine-tuning (Wu et al., 2024). Anatomy-aware unrolling uses segmentation-derived target locations to guide multi-contrast MRI reconstruction (He et al., 2025), while tunable CT networks optimise image quality for lesion detectability (Tivnan et al., 2023). These methods introduce task information through different mechanisms, including acquisition design, diagnostic losses and anatomical supervision.

Evaluation must distinguish fidelity, visual realism and diagnostic utility. The perception–distortion tradeoff separates perceptual quality from distortion (Blau & Michaeli, 2018), without establishing a corresponding diagnostic tradeoff. More directly, human MRI observers reveal limitations of global metrics for subtle-signal detection (Herman et al., 2024), and AI-observer studies assess sparse-view CT through haemorrhage diagnosis (Tivnan et al., 2025). CSF-Recon uses a learned spatial field to connect restoration with diagnostic pooling under prescribed acquisitions. The probes in Appendix C.9 additionally distinguish reconstruction readability from diagnostic-head contributions.

B.3 Attribution reliability and explanation-constrained learning

Grad-CAM uses class gradients to localise responses in convolutional feature maps (Selvaraju et al., 2017); Integrated Gradients accumulates input gradients relative to a baseline (Sundararajan et al., 2017). Plausible maps nevertheless require validation. Randomisation tests reveal that some saliency methods are insensitive to learned model parameters or training labels (Adebayo et al., 2018), while medical-imaging evaluations identify limitations in localisation, repeatability and reproducibility (Arun et al., 2021). Radiographic COVID-19 classifiers further demonstrate that accurate prediction can rely on shortcuts (DeGrave et al., 2021).

Explanation-constrained learning addresses such dependencies during training. Right for the Right Reasons penalises input gradients in irrelevant regions (Ross et al., 2017); contextual decomposition explanation penalisation constrains feature importance and interactions using prior knowledge (Rieger et al., 2020). CSF-Recon similarly seeks informative spatial guidance, but tests its utility through restoration and diagnosis. Lesion-restricted teacher responses supervise a predicted support field; their appearance alone is not taken as evidence of faithful segmentation or causality.

B.4 Counterfactual alignment and relation to CSF-Recon

Counterfactual explanations identify input changes that alter a model decision (Wachter et al., 2018), while contrastive explanations distinguish sufficient present evidence from decision-changing absent features (Dhurandhar et al., 2018). In medical imaging, Singla et al. generate gradual chest-radiograph transformations that change classifier outputs (Singla et al., 2023). Attri-Net bases multi-label predictions on class-specific counterfactual attribution maps (Sun et al., 2024). These explana-

tions depend on the predictor and allowed perturbations, and do not by themselves identify disease mechanisms.

Liu et al. extend counterfactual reasoning to training by aligning decision-changing evidence with expert annotations through an implicitly differentiated objective (Liu et al., 2025). CSF-Recon adopts this principle for a reference teacher and transfers its lesion-restricted class responses to a student operating on incomplete measurements. The predicted field then gates high-frequency injection, local refinement and diagnostic pooling while retaining global context. Its contribution is the use of aligned evidence for task-preserving reconstruction, rather than a new counterfactual criterion. Teacher supervision and annotations are required only during training, and the field's effectiveness is assessed through reconstruction, ablation and diagnostic-probe experiments.

C Implementation Details and Additional Reporting

C.1 Acquisition and reconstruction implementation

The golden-angle MR masks are rasterised on a Cartesian k-space grid and reconstructed with FFT/SENSE rather than a continuous-trajectory NUFFT. Measured multicoil k-space is retained. ESPIRiT sensitivity maps are estimated from a central calibration width of 24, and the coil dimension is padded to 24. The fully sampled MR reference is formed by the root-sum-of-squares (RSS) magnitude of the coil images, $\mathbf{x}_{\mathrm{gt}} = \left(\sum_{c=1}^{C} |\mathcal{F}^{-1}\mathbf{b}_c^{\mathrm{full}}|^2\right)^{1/2}$, followed by normalisation to $[0, 1]$. CT measurements are simulated from complete 720-view parallel-beam projections and reduced to 45 views for the main $R = 16$ setting. All methods use identical preprocessing, patient- or volume-level splits, measurements and fixed test operators.

C.2 ISTA-Net+ coarse reconstructor

ISTA-Net+ interprets iterative shrinkage-thresholding as a finite-depth network (Zhang & Ghanem, 2018). Each phase retains a physics-based data-consistency step and replaces the analytic proximal operator with a learned nonlinear mapping. In a modality-independent form, phase k is

$$\mathbf{r}^{(k)} = \mathbf{x}^{(k-1)} - \rho_k \mathcal{A}_d^H \left(\mathcal{A}_d \mathbf{x}^{(k-1)} - \mathbf{b}_d\right), \qquad \mathbf{x}^{(k)} = \mathbf{r}^{(k)} + \mathcal{P}_{\phi_k}\left(\mathbf{r}^{(k)}\right), \tag{22}$$

where ρ_k is a learned step size and $\mathcal{P}_{\phi_k}$ is the learned residual proximal mapping. Thus, measurement consistency is imposed through $\mathcal{A}_d$ and $\mathcal{A}_d^H$, while the proximal branch learns the image prior.

Our coarse reconstructor contains nine unrolled phases, each comprising a data-consistency block, a learned proximal block and a residual connection (Figure 7). For MR, the multicoil adjoint provides a two-channel real–imaginary initialisation at 320×320, and the final iterate is converted to a clipped magnitude image. For CT, FBP initialises a single-channel 96×96 image and the output is clipped to $[0, 1]$. The network is pretrained with reconstruction supervision and then frozen; its output $\mathbf{x}_c$ serves as the common coarse input to CSF-Recon.

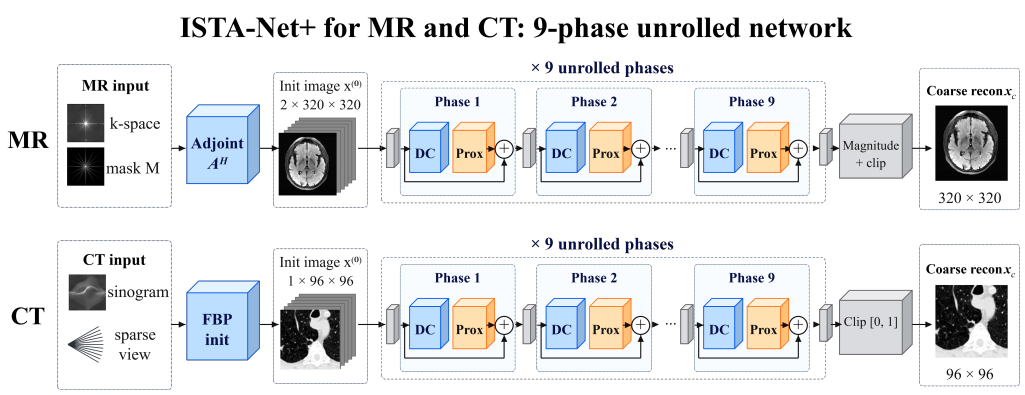

Figure 7: Modality-specific ISTA-Net+ coarse reconstruction. MR uses multicoil adjoint initialisation and magnitude output, whereas CT uses FBP initialisation; both paths share the same nine-phase data-consistency and learned-proximal design.

C.3 Local residual refiner

A single microstructure-injection step may not fully correct lesion-region errors inherited from $\mathbf{x}_c$. We therefore apply a lightweight iterative refiner after $\hat{\mathbf{x}}_0$, using the predicted support field $\mathbf{m}$ directly as the spatial gate. With $\mathbf{x}^{(0)} = \hat{\mathbf{x}}_0$, step t predicts

$$\boldsymbol{\delta}_t = \mathcal{F}_{\text{ref}}\Big(\mathbf{x}^{(t)}, \mathbf{x}_c, \mathbf{x}_{\text{base}}, \mathbf{x}_{\text{micro}}, \mathbf{m}, \mathcal{P}(\mathbf{f}_c), t/T\Big), \tag{23}$$

and updates

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \mathbf{m} \odot \boldsymbol{\delta}_t, \qquad \hat{\mathbf{x}}_T = \mathbf{x}^{(T)}. \tag{24}$$

The refiner is trained separately with the target support field $\mathbf{m}^\star$ weighting the reconstruction error:

$$\mathcal{L}_{\text{ref}} = \frac{\|\mathbf{m}^\star \odot (\hat{\mathbf{x}}_T - \mathbf{x}_{\text{gt}})\|_2^2}{\|\mathbf{m}^\star\|_1 + \epsilon}. \tag{25}$$

The current configuration performs six iterative correction steps with shared refiner parameters. At each step, the current reconstruction, coarse and structured components, support field, compressed encoder features and broadcast normalised step index are concatenated into a 38-channel tensor. As shown in Figure 8, a 3×3 stem maps this tensor to 64 channels, six residual blocks process the features, and a 3×3 prediction head with $\tanh$ and a scale factor of 0.5 produces the single-channel residual $\boldsymbol{\delta}_t$. Restricting each update with $\mathbf{m}$ concentrates refinement on diagnostically relevant regions and limits unnecessary changes elsewhere.

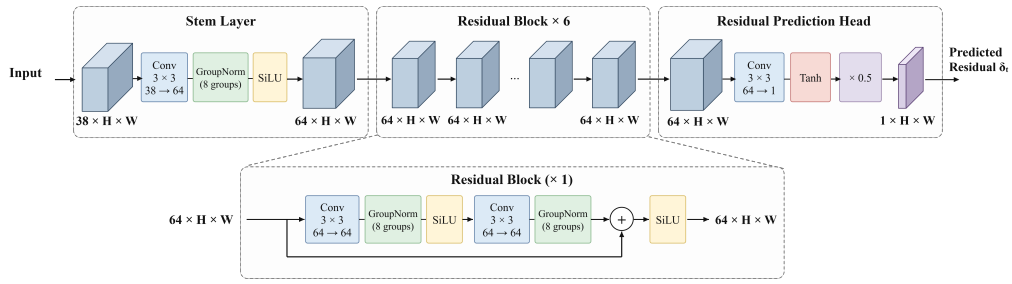

Figure 8: Architecture of the shared local residual refiner used at each correction step. A 38-channel input is projected to 64 channels, processed by six residual blocks, and mapped to the bounded single-channel residual $\boldsymbol{\delta}_t$.

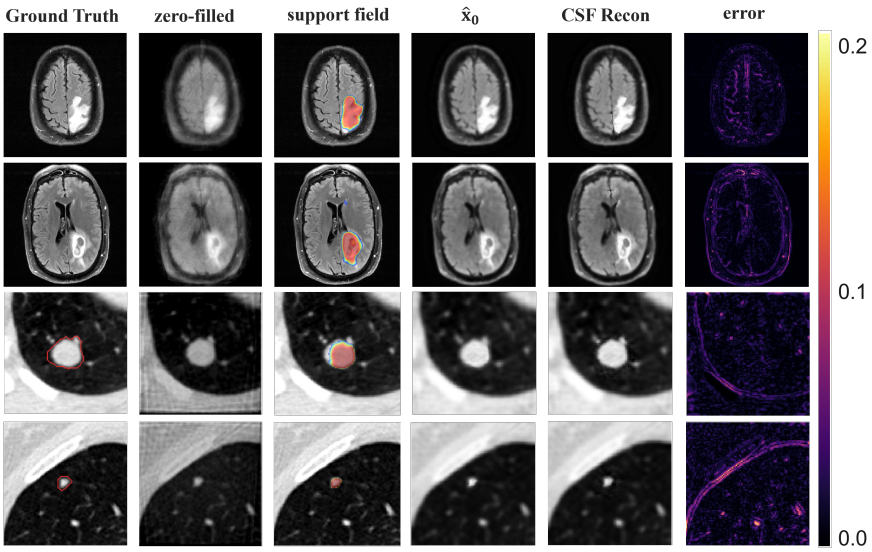

Figure 9: Component visualisation of CSF-Recon. The predicted support field identifies diagnostically relevant regions, the support-gated initial reconstruction $\hat{x}_0$ restores candidate local structure, and the final CSF-Recon output further refines lesion-region detail.

C.4 Global fidelity versus diagnostic utility

Figure 10 complements the ROI analysis in Figure 4 by replacing ROI-PSNR with whole-image PSNR. The same mismatch remains: methods with higher global pixel fidelity do not necessarily provide greater diagnostic utility. Tables 4 and 5 report the complementary pixel-fidelity metrics for MR and CT, respectively.

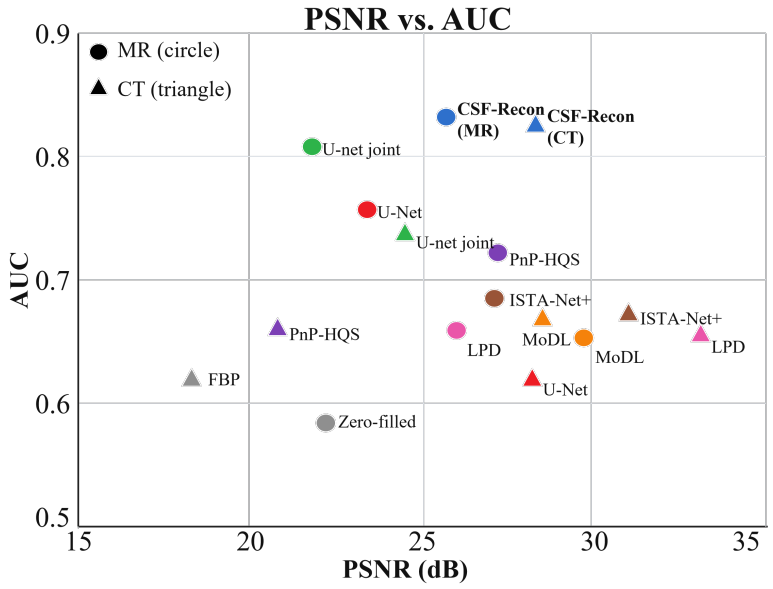

Figure 10: Whole-image fidelity versus diagnostic utility. Circles and triangles denote MR and CT, respectively. The methods with the highest PSNR do not necessarily attain the highest AUC.

Table 4: Complementary reconstruction-fidelity results on MR fastMRI+. Results are mean ± standard deviation over three random seeds when available. Bold and underlined values indicate the best and second-best means, respectively.

Method	PSNR ↑	SSIM ↑
Zero-filled	21.58 ± 0.00	0.9192 ± 0.0000
U-Net	23.52 ± 0.17	0.9557 ± 0.0024
U-Net joint	22.39 ± 0.23	0.9295 ± 0.0059
LPD	25.53 ± 0.04	0.9706 ± 0.0004
PnP-HQS	27.23 ± 0.00	0.9780 ± 0.0000
MoDL	$\underline{27.28 \pm 0.22}$	$\underline{0.9812 \pm 0.0009}$
ISTA-Net+	$\mathbf{27.29 \pm 0.00}$	$\mathbf{0.9818 \pm 0.0000}$
CSF-Recon	26.09 ± 0.09	0.9642 ± 0.0008

Table 5: Complementary reconstruction-fidelity results on CT LIDC-IDRI. Results are mean ± standard deviation over three random seeds when available. Bold and underlined values indicate the best and second-best means, respectively.

Method	PSNR ↑	SSIM ↑
Zero-filled / FBP	18.73 ± 0.00	0.8813 ± 0.0000
U-Net	27.13 ± 0.50	0.9774 ± 0.0019
U-Net joint	25.61 ± 1.12	0.9635 ± 0.0097
LPD	$\mathbf{32.22 \pm 0.14}$	$\mathbf{0.9885 \pm 0.0002}$
PnP-HQS	20.49 ± 0.00	0.9097 ± 0.0000
MoDL	29.25 ± 0.12	0.9841 ± 0.0003
ISTA-Net+	$\underline{30.56 \pm 0.00}$	$\underline{0.9872 \pm 0.0000}$
CSF-Recon	28.13 ± 0.03	0.9729 ± 0.0002

C.5 Paired case-level analysis of diagnostic accuracy

Although the main metrics in Table 1 are computed at the slice level, slices or lesion patches from the same case are not statistically independent. We therefore assessed the robustness of the diagnostic-accuracy gains using existing test predictions without retraining. MR volumes and CT patients were treated as the independent statistical units, respectively; case identifiers, labels and random-seed settings were strictly matched between CSF-Recon and each baseline.

We performed 5,000 paired bootstrap resamples with replacement. Within each replicate, the same sampled cases were used for both methods and all three seeds. Accuracy was computed separately for each seed and then averaged across seeds. For each baseline, the paired difference was

$$\Delta\mathrm{Acc} = \mathrm{Acc}_{\text{CSF-Recon}} - \mathrm{Acc}_{\text{baseline}}, \tag{26}$$

so positive values favour CSF-Recon. The 95% confidence interval was defined by the 2.5th and 97.5th percentiles of the bootstrap distribution. The two-sided paired-bootstrap p-value was computed as twice the smaller proportion of bootstrap differences on either side of zero.

Table 6: Paired case-level analysis of diagnostic accuracy. ΔAcc. denotes CSF-Recon minus the corresponding baseline. MR volumes and CT patients are the independent units. Confidence intervals and two-sided p-values are estimated from 5,000 paired bootstrap resamples over three random seeds.

	MR			CT		
Baseline	ΔAcc.	95% CI	Bootstrap p	ΔAcc.	95% CI	Bootstrap p
Zero-filled / FBP	$+0.253$	$[0.107, 0.400]$	0.003	$+0.107$	$[0.037, 0.181]$	0.005
U-Net	$+0.213$	$[0.027, 0.400]$	0.035	$+0.156$	$[0.078, 0.233]$	< 0.001
U-Net joint	$+0.112$	$[0.093, 0.182]$	0.043	$+0.076$	$[0.068, 0.092]$	0.006
LPD	$+0.253$	$[0.053, 0.453]$	0.018	$+0.115$	$[0.037, 0.196]$	0.007
PnP-HQS	$+0.253$	$[0.053, 0.440]$	0.018	$+0.104$	$[0.022, 0.181]$	0.014
MoDL	$+0.267$	$[0.040, 0.480]$	0.023	$+0.126$	$[0.044, 0.211]$	0.003
ISTA-Net+	$+0.147$	$[0.013, 0.307]$	0.041	$+0.104$	$[0.022, 0.185]$	0.017

All paired differences are positive, all 95% confidence intervals exclude zero, and all two-sided p-values are below 0.05. Thus, the diagnostic-accuracy advantage of CSF-Recon remains reliable when dependence among slices from the same MR volume or CT patient is accounted for.

C.6 Additional CT support-field visualisation

Figure 11 extends the support-field source ablation in Section 3.3 with additional CT cases.

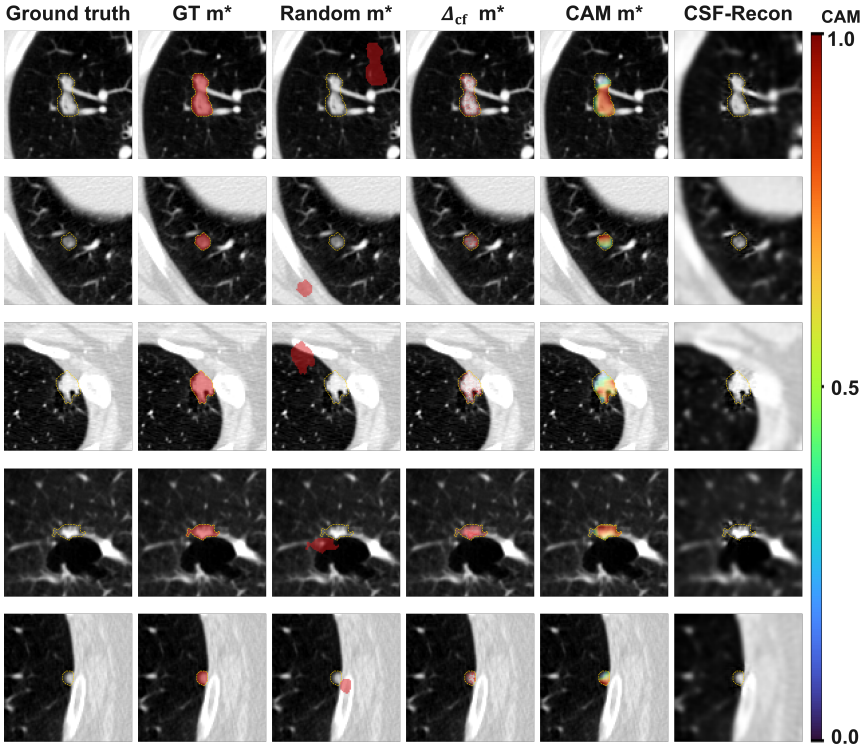

Figure 11: Additional CT support-field source comparisons. Columns show the ground-truth image, support fields obtained from direct lesion-annotation supervision, a randomly shifted annotation, the annotation-restricted counterfactual difference, and the annotation-restricted CAM, followed by the corresponding CSF-Recon output. Yellow dashed contours indicate lesion annotations; coloured overlays denote support fields.

C.7 PLAIN VERSUS COUNTERFACTUALLY ALIGNED GRAD-CAM

To isolate the effect of counterfactual alignment, the plain and aligned teachers use the same classifier architecture, training data and classification objective. The plain classifier is updated only by classification loss, whereas the aligned teacher additionally receives $\mathcal{L}_{\text{align}}$. Their true-class Grad-CAMs are computed and used to supervise the student through the same support-field pipeline, so the comparison changes only the teacher-alignment objective.

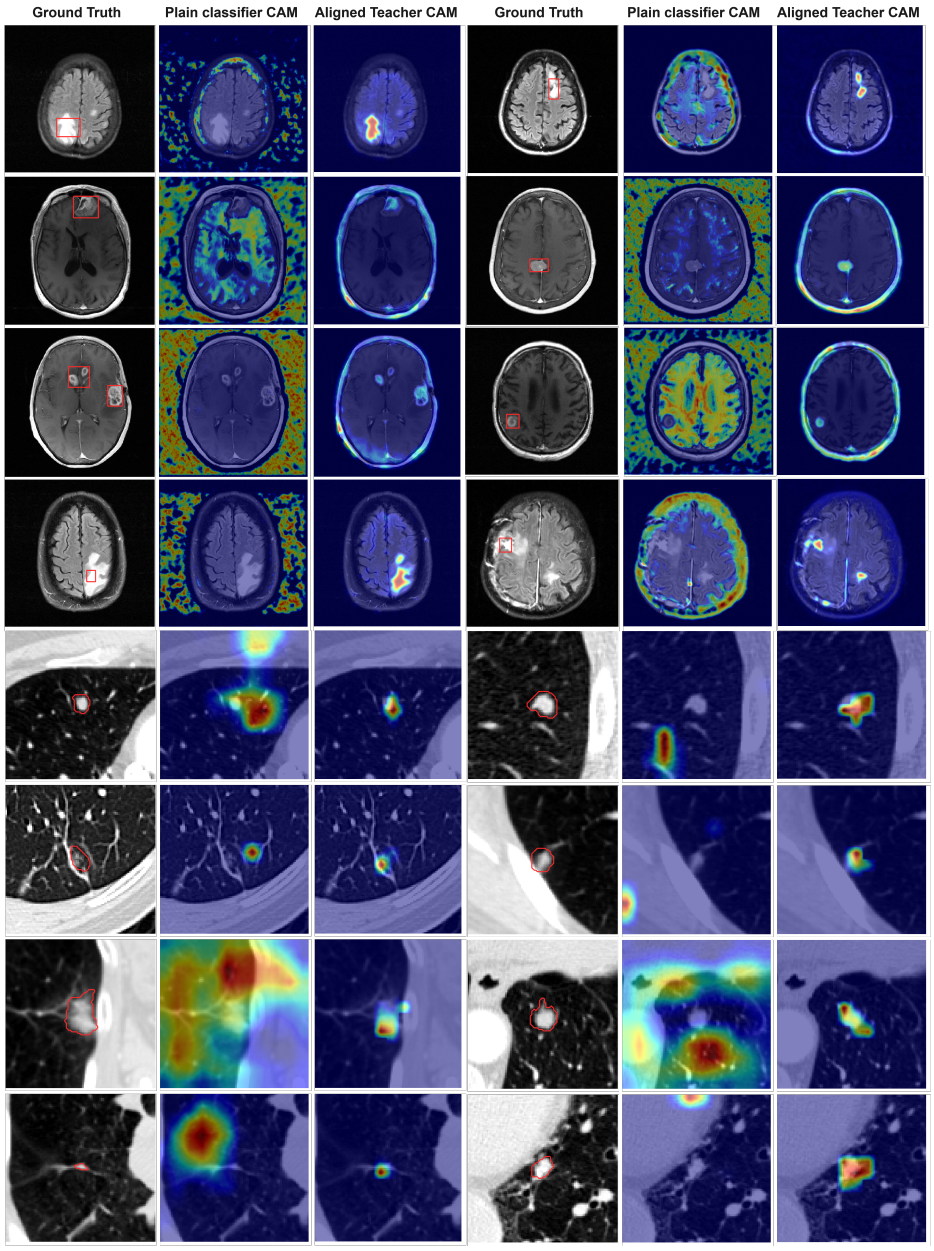

Figure 12: Grad-CAM comparison between the classification-only plain teacher and the counterfactually aligned teacher. Counterfactual alignment suppresses extra-lesional responses and concentrates class-relevant evidence more coherently around the annotated lesion, providing a more suitable reconstruction support target.

The aligned maps are more lesion-focused and spatially coherent than the plain-classifier maps. The corresponding gains over Plain Grad-CAM in Table 3 show that this improvement is not merely

visual: the aligned attention provides more effective supervision for both local reconstruction and downstream diagnosis.

C.8 Decision-Sensitivity Alignment of the Learned Support Field

To connect the analysis in Section 2.7 with the operational support field without circularity, we use a separately trained and frozen diagnostic probe (fixed seed 2026). The probe shares no parameters with the CSF-Recon teacher or student and receives no support-field supervision. For each fully sampled test image $\mathbf{x}_{\mathrm{gt}}$, we define the independent pixel-wise decision-sensitivity reference as

$$\mathbf{s}(\mathbf{x}_{\mathrm{gt}}) = \left|\nabla_{\mathbf{x}_{\mathrm{gt}}} \left[z_y(\mathbf{x}_{\mathrm{gt}}) - z_{1-y}(\mathbf{x}_{\mathrm{gt}})\right]\right|, \tag{27}$$

where z_y and z_{1-y} are the probe logits for the ground-truth and alternative classes. Each sensitivity map is normalised sample-wise to $[0, 1]$.

We compare $\mathbf{s}$ with the true-class Grad-CAM of the classification-only teacher ($\mathbf{M}_{\mathrm{plain}}$), the Grad-CAM of the counterfactually aligned teacher ($\mathbf{M}_{\mathrm{align}}$), and the student-predicted inference-time support field ($\mathbf{M}_{\mathrm{pred}}$). None of these maps is multiplied by a lesion mask for this analysis; all are resized to the resolution of $\mathbf{s}$ and normalised sample-wise to $[0, 1]$. Spatial agreement is measured per test image using

$$\rho = \mathrm{Spearman}(\mathrm{vec}(\mathbf{s}), \mathrm{vec}(\mathbf{M})), \qquad O_{10} = \frac{|\,\mathrm{Top}_{10\%}(\mathbf{s}) \cap \mathrm{Top}_{10\%}(\mathbf{M})|}{|\,\mathrm{Top}_{10\%}(\mathbf{s})|}. \tag{28}$$

The first metric compares whole-image pixel rankings, whereas O_{10} measures overlap between the highest-response regions. Independent random top-10% selections have expected overlap 0.10. Table 7 reports test-set means.

Table 7: Spatial agreement with the independent decision-sensitivity reference. Higher values indicate stronger agreement.

Modality	Spatial map	Spearman ↑	Top-10% overlap ↑
MR	$\mathbf{M}_{\mathrm{plain}}$	-0.0761	0.1517
	$\mathbf{M}_{\mathrm{align}}$	$\underline{0.2076}$	$\underline{0.2432}$
	$\mathbf{M}_{\mathrm{pred}}$	$\mathbf{0.4596}$	$\mathbf{0.3357}$
CT	$\mathbf{M}_{\mathrm{plain}}$	0.1140	0.1483
	$\mathbf{M}_{\mathrm{align}}$	$\underline{0.1790}$	$\underline{0.2356}$
	$\mathbf{M}_{\mathrm{pred}}$	$\mathbf{0.2086}$	$\mathbf{0.3089}$

Counterfactual alignment increases both measures relative to Plain Grad-CAM in MR and CT. The student-predicted field attains the strongest agreement in both modalities, giving the same ordering, $\mathbf{M}_{\mathrm{pred}} > \mathbf{M}_{\mathrm{align}} > \mathbf{M}_{\mathrm{plain}}$, for both metrics. Thus, the aligned teacher CAM shows stronger spatial correspondence with independently measured decision sensitivity, and the learned student field preserves and, in these experiments, further consolidates this structure. These results provide an empirical bridge between the decision-sensitive reconstruction analysis and the operational support-field mechanism. They support the learned field as an empirical spatial proxy for decision-sensitive structure; they do not imply that it is identical to input-gradient sensitivity or constitutes a causal attribution.

C.9 Diagnostic utility of the final reconstruction

We use two complementary probe protocols to determine whether reconstruction gains correspond to usable diagnostic information. The *fixed probe* is trained only on fully sampled $\mathbf{x}_{\mathrm{gt}}$, frozen, and then applied without adaptation to zero-filled/FBP images, $\mathbf{x}_c$, $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{x}}_T$. It therefore measures how well each reconstruction preserves evidence in the diagnostic representation learned from the reference domain. The *matched probes* are trained independently for each representation with identical architectures, patient-level splits and optimisation settings. During this comparison, the reconstruction network, shared encoder, support predictor and refiner remain frozen, so differences reflect the diagnostic information available from each representation and pooling strategy rather than changes to reconstruction.

Table 8: MR diagnostic probe experiments on fastMRI+. Panel (a) uses a frozen probe trained on fully sampled images; panel (b) uses independently trained matched probes.

(a) Fixed probe

Metric	Oracle $\mathbf{x}_{\mathrm{gt}}$	ZF	$\mathbf{x}_c$	$\hat{\mathbf{x}}_0$	$\hat{\mathbf{x}}_T$
AUC ↑	**0.945**	0.606	0.805	0.803	0.875
Acc. ↑	**0.847**	0.726	0.798	0.766	0.815
Bal. Acc. ↑	**0.748**	0.500	0.671	0.583	0.709
F1 ↑	**0.655**	—	0.511	0.393	0.587

(b) Matched training probes

Metric	$\mathbf{x}_c$ Global	$\hat{\mathbf{x}}_0$ Global	$\hat{\mathbf{x}}_T$ CNN	$\hat{\mathbf{x}}_T$ Global	$\hat{\mathbf{x}}_T$ Local	$\hat{\mathbf{x}}_T$ L+G
AUC ↑	0.860	0.825	0.845	0.833	0.833	**0.871**
Acc. ↑	0.637	0.750	0.734	0.766	0.718	**0.798**
F1 ↑	0.587	0.597	0.561	0.603	0.607	**0.621**

Table 9: CT diagnostic probe experiments on LIDC-IDRI, reported as mean ± standard deviation over three random seeds. In panel (a), bold marks the strongest reconstructed input, while the oracle is the fully sampled reference. Panel (b) compares independently trained matched probes with frozen reconstruction components.

(a) Fixed probe

Metric	Oracle $\mathbf{x}_{\mathrm{gt}}$	FBP	$\mathbf{x}_c$	$\hat{\mathbf{x}}_0$	$\hat{\mathbf{x}}_T$
AUC ↑	0.733 ± 0.009	0.704 ± 0.012	0.718 ± 0.008	0.720 ± 0.008	$\mathbf{0.728 \pm 0.008}$
Acc. ↑	0.678 ± 0.005	0.650 ± 0.027	0.656 ± 0.006	0.671 ± 0.020	$\mathbf{0.672 \pm 0.007}$
Bal. Acc. ↑	0.681 ± 0.003	0.651 ± 0.022	0.670 ± 0.010	0.675 ± 0.015	$\mathbf{0.676 \pm 0.002}$
F1 ↑	0.681 ± 0.021	0.661 ± 0.046	0.667 ± 0.012	0.672 ± 0.041	$\mathbf{0.674 \pm 0.026}$

(b) Matched training probes

Metric	$\mathbf{x}_c$ Global	$\hat{\mathbf{x}}_0$ Global	$\hat{\mathbf{x}}_T$ CNN	$\hat{\mathbf{x}}_T$ Global	$\hat{\mathbf{x}}_T$ Local	$\hat{\mathbf{x}}_T$ L+G
AUC ↑	0.811 ± 0.010	0.811 ± 0.013	0.757 ± 0.002	0.811 ± 0.012	0.826 ± 0.021	$\mathbf{0.828 \pm 0.016}$
Acc. ↑	0.717 ± 0.047	0.730 ± 0.030	0.690 ± 0.022	0.735 ± 0.010	0.690 ± 0.089	$\mathbf{0.747 \pm 0.010}$
Bal. Acc. ↑	0.714 ± 0.043	0.729 ± 0.026	0.690 ± 0.013	0.724 ± 0.020	0.669 ± 0.102	$\mathbf{0.744 \pm 0.017}$
F1 ↑	0.735 ± 0.074	0.746 ± 0.047	0.703 ± 0.049	0.767 ± 0.038	0.769 ± 0.012	$\mathbf{0.775 \pm 0.014}$

On CT, the fixed probe consistently favours $\hat{\mathbf{x}}_T$ among the reconstructed inputs: its AUC increases from 0.718 for $\mathbf{x}_c$ and 0.720 for $\hat{\mathbf{x}}_0$ to 0.728, approaching the oracle AUC of 0.733. This indicates that refinement restores evidence that remains directly readable by a classifier trained only on reference images, rather than merely enabling classifier adaptation. Under matched training, local–global fusion achieves the best CT result for every metric (AUC 0.828, accuracy 0.747, balanced accuracy 0.744 and F1 0.775), outperforming either pooling branch and the conventional CNN probe. The same overall tendency is observed for MR: the final reconstruction is preferred by the fixed probe, while fused local and global evidence performs best under matched training. These two protocols jointly show that CSF-Recon both preserves reference-domain diagnostic cues and provides complementary local and global information for an adapted diagnostic model.

An ROI-PSNR gain without a corresponding fixed- or matched-probe AUC gain would instead indicate pixel fitting rather than restoration of diagnostic microstructure.

Fixed probe across reconstruction methods. To distinguish reconstruction gains from diagnostic-head design, we use a common frozen probe within each modality. Its classifier-head architecture is the same as that used for the baselines and U-Net joint, but differs from the within-method fixed probe above. The probe is trained only on fully sampled reference images, then frozen

and applied to each reconstructed input without method-specific adaptation or support-weighted pooling.

Table 10: Cross-method frozen-probe comparison. Slice-level accuracy and AUC are reported as mean ± standard deviation. Bold denotes the highest mean among reconstructed inputs; fully sampled references are shown separately.

	MR (fastMRI+)		CT (LIDC-IDRI)	
Input image	Accuracy	AUC	Accuracy	AUC
Fully sampled reference	0.730 ± 0.056	0.802 ± 0.084	0.609 ± 0.005	0.679 ± 0.008
Zero-filled / FBP	0.650 ± 0.042	0.715 ± 0.037	0.565 ± 0.035	0.625 ± 0.016
U-Net	0.634 ± 0.055	0.779 ± 0.048	0.580 ± 0.004	0.648 ± 0.019
LPD	0.676 ± 0.040	0.679 ± 0.032	0.573 ± 0.017	0.652 ± 0.018
MoDL	0.635 ± 0.073	0.695 ± 0.128	0.578 ± 0.013	0.648 ± 0.013
ISTA-Net+	0.670 ± 0.072	0.733 ± 0.100	0.573 ± 0.019	0.653 ± 0.015
U-Net joint	0.654 ± 0.067	0.756 ± 0.092	0.580 ± 0.026	0.667 ± 0.014
CSF-Recon	$\mathbf{0.686 \pm 0.019}$	$\mathbf{0.792 \pm 0.022}$	$\mathbf{0.606 \pm 0.017}$	$\mathbf{0.673 \pm 0.019}$

Table 10 shows that CSF-Recon achieves the highest mean accuracy and AUC among reconstructed inputs in both modalities. Its AUC reaches 0.792 for MR and 0.673 for CT, supporting a contribution from the reconstructed images themselves beyond CSF-Recon's diagnostic-head design.

C.10 Rationale for support-gated frequency allocation

Task-adapted reconstruction couples image recovery to downstream objectives (Adler et al., 2022), but an image-level diagnostic loss alone does not specify where limited recovery capacity should be allocated. CSF-Recon therefore adopts the design approximation

$$\mathcal{D}(\mathbf{x}) \approx \mathcal{D}(\mathcal{L}(\mathbf{x}), \mathbf{m} \odot \mathcal{H}(\mathbf{x})), \tag{29}$$

where $\mathcal{L}$ and $\mathcal{H}$ are the low- and high-frequency operators and $\mathbf{m} \in [0,1]^{H \times W}$ is the predicted support field. This approximation expresses an architectural decomposition: global low-frequency anatomy supplies contextual information, while the support field selects high-frequency evidence for preferential recovery. It does not assert that the selected region is sufficient for diagnosis or that $\mathbf{m}$ is a causal attribution map.

Radial MR and sparse-view CT errors contain structured high-frequency streaks. Restoring high-frequency content uniformly can recover anatomical detail but can also reintroduce measurement-induced artefacts. The base branch therefore preserves low-frequency anatomy over the full image, whereas $\mathbf{m}$ gates microstructure injection and residual refinement. The diagnostic head follows the same principle by combining support-weighted local pooling with global pooling, so extra-lesional context is retained rather than discarded.

The approximation is evaluated empirically rather than treated as a guarantee. The component and support-source ablations in Section 3.3 test whether spatial gating, refinement and the source of $\mathbf{m}$ affect reconstruction and diagnosis, while the fixed and matched probes in Appendix C.9 examine whether the final reconstruction preserves reference-readable diagnostic evidence. These controls bound the interpretation of Equation (29) to a testable design rationale.

D Analysis of Decision-Sensitive Reconstruction

D.1 Exact path-wise decision-preservation bound

Let $\mathbf{e} = \hat{\mathbf{x}} - \mathbf{x}^\star$ and define

$$h(t) = \mathcal{G}(\mathbf{x}^\star + t\mathbf{e}), \qquad t \in [0,1]. \tag{30}$$

Assume that h is absolutely continuous on $[0,1]$, that $\mathcal{G}$ is differentiable along this path almost everywhere, that

$$h'(t) = \nabla \mathcal{G}(\mathbf{x}^\star + t\mathbf{e})^\top \mathbf{e} \tag{31}$$

holds almost everywhere, and that the pathwise partial derivatives $t \mapsto \partial_i \mathcal{G}(\mathbf{x}^\star + t\mathbf{e})$ are integrable. By absolute continuity,

$$h(1) - h(0) = \int_0^1 h'(t)\, dt. \tag{32}$$

Therefore,

$$\mathcal{G}(\hat{\mathbf{x}}) - \mathcal{G}(\mathbf{x}^\star) = \int_0^1 \nabla \mathcal{G}(\mathbf{x}^\star + t\mathbf{e})^\top \mathbf{e}\, dt = \sum_i e_i \int_0^1 \partial_i \mathcal{G}(\mathbf{x}^\star + t\mathbf{e})\, dt. \tag{33}$$

Using $ab \geq -|a||b|$ and the integral triangle inequality,

$$\begin{aligned} \mathcal{G}(\hat{\mathbf{x}}) - \mathcal{G}(\mathbf{x}^\star) &\geq -\sum_i |e_i| \left| \int_0^1 \partial_i \mathcal{G}(\mathbf{x}^\star + t\mathbf{e})\, dt \right| \\ &\geq -\sum_i |e_i| \int_0^1 |\partial_i \mathcal{G}(\mathbf{x}^\star + t\mathbf{e})|\, dt. \end{aligned} \tag{34}$$

With

$$\bar{s}_i(\mathbf{e}) = \int_0^1 |\partial_i \mathcal{G}(\mathbf{x}^\star + t\mathbf{e})|\, dt, \tag{35}$$

we obtain

$$\mathcal{G}(\hat{\mathbf{x}}) \geq \mathcal{G}(\mathbf{x}^\star) - \sum_i \bar{s}_i(\mathbf{e}) |e_i|. \tag{36}$$

If $\mathcal{G}(\mathbf{x}^\star) > 0$ and $\sum_i \bar{s}_i(\mathbf{e})|e_i| < \mathcal{G}(\mathbf{x}^\star)$, then Equation (36) gives $\mathcal{G}(\hat{\mathbf{x}}) > 0$, so the binary reference decision is preserved.

Remark D.1. Equation (36) is an exact path-wise inequality. It does not require a globally Lipschitz gradient or a second-order Taylor approximation. It also does not imply that sufficiently small global MSE cannot preserve a decision; instead, it exposes a spatial sensitivity weighting that conventional global fidelity measures do not explicitly encode. For sufficiently local reconstruction perturbations, if the diagnostic gradient varies little along the path,

$$\bar{s}_i(\mathbf{e}) \approx |\partial_i \mathcal{G}(\mathbf{x}^\star)|. \tag{37}$$

D.2 Local geometry of the quadratic counterfactual surrogate

For the fixed binary diagnostic classifier, $\mathcal{G}(\mathbf{x}) = \mathcal{F}_y(\mathbf{x}) - \mathcal{F}_{\bar{y}}(\mathbf{x})$. Under standard unweighted binary softmax cross-entropy,

$$p_{\bar{y}}(\mathbf{x}) = \frac{\exp(\mathcal{F}_{\bar{y}}(\mathbf{x}))}{\exp(\mathcal{F}_y(\mathbf{x})) + \exp(\mathcal{F}_{\bar{y}}(\mathbf{x}))} = \frac{1}{1 + \exp(\mathcal{G}(\mathbf{x}))}, \tag{38}$$

and hence

$$-\log p_{\bar{y}}(\mathbf{x}) = \log(1 + \exp(\mathcal{G}(\mathbf{x}))). \tag{39}$$

For $\mathbf{x} = \mathbf{x}^\star + \boldsymbol{\delta}$, locally linearize

$$\mathcal{G}(\mathbf{x}^\star + \boldsymbol{\delta}) \approx G_0 + \mathbf{g}^\top \boldsymbol{\delta}, \qquad G_0 = \mathcal{G}(\mathbf{x}^\star), \quad \mathbf{g} = \nabla \mathcal{G}(\mathbf{x}^\star). \tag{40}$$

The derivation below assumes that $\boldsymbol{\delta}$ is unconstrained, standard unweighted binary softmax cross-entropy is used, the proximity term is

$$\mathcal{D}(\mathbf{x}^\star + \boldsymbol{\delta}, \mathbf{x}^\star) = \frac{1}{2}\|\boldsymbol{\delta}\|_2^2, \tag{41}$$

and there is no clipping, box constraint, or additional regularizer. The local surrogate is

$$\mathcal{L}_{\text{sur}}(\boldsymbol{\delta}) = \log\big(1 + \exp(G_0 + \mathbf{g}^\top \boldsymbol{\delta})\big) + \frac{\lambda_{\text{cf}}}{2}\|\boldsymbol{\delta}\|_2^2. \tag{42}$$

Let $z = G_0 + \mathbf{g}^\top \boldsymbol{\delta}$. Its gradient and Hessian are

$$\nabla_{\boldsymbol{\delta}} \mathcal{L}_{\text{sur}} = \sigma(z)\mathbf{g} + \lambda_{\text{cf}} \boldsymbol{\delta}, \tag{43}$$

$$\nabla^2_{\boldsymbol{\delta}} \mathcal{L}_{\text{sur}} = \sigma(z)\big(1 - \sigma(z)\big)\mathbf{g}\mathbf{g}^\top + \lambda_{\text{cf}} \mathbf{I}, \tag{44}$$

where $\sigma(z) = (1 + \mathrm{e}^{-z})^{-1}$. Since $\sigma(z)(1 - \sigma(z))\mathbf{g}\mathbf{g}^\top \succeq 0$ and $\lambda_{\mathrm{cf}} > 0$,

$$\nabla^2_{\boldsymbol{\delta}} \mathcal{L}_{\mathrm{sur}} \succeq \lambda_{\mathrm{cf}} \mathbf{I} \succ 0. \tag{45}$$

Thus, $\mathcal{L}_{\mathrm{sur}}$ is λ_{cf}-strongly convex and has a unique global minimizer $\boldsymbol{\delta}^\star_{\mathrm{sur}}$. At the minimizer,

$$\lambda_{\mathrm{cf}} \boldsymbol{\delta}^\star_{\mathrm{sur}} + \sigma(z^\star)\mathbf{g} = 0, \tag{46}$$

so

$$\boldsymbol{\delta}^\star_{\mathrm{sur}} = -\frac{\sigma(z^\star)}{\lambda_{\mathrm{cf}}}\mathbf{g} = -c\nabla\mathcal{G}(\mathbf{x}^\star), \qquad c = \frac{\sigma(z^\star)}{\lambda_{\mathrm{cf}}} > 0. \tag{47}$$

Consequently,

$$|\delta^\star_{\mathrm{sur},i}| = c|\partial_i \mathcal{G}(\mathbf{x}^\star)|. \tag{48}$$

Thus, under the stated assumptions, the spatial magnitude pattern of the local surrogate minimizer is proportional to diagnostic-margin sensitivity. Equation (47) is not a general analytical characterization of a finite-step or early-stopped counterfactual trajectory.

Remark D.2: normalization of the quadratic term. The coefficient λ_{cf} above corresponds to the quadratic term $\frac{1}{2}\|\boldsymbol{\delta}\|_2^2$. If the implementation instead uses $\frac{1}{2n_{\mathrm{pix}}}\|\boldsymbol{\delta}\|_2^2$, the same derivation holds with the effective coefficient

$$\lambda_{\mathrm{eff}} = \frac{\lambda_{\mathrm{cf}}}{n_{\mathrm{pix}}}. \tag{49}$$

More generally, the sum/mean reduction used in the implementation must be absorbed consistently into the effective quadratic coefficient.

Remark D.3: scope of the analytical result. If the implementation uses clipping, box constraints, TV, SSIM, perceptual penalties, spatially weighted quadratic penalties, or other regularizers or constraints, Equation (47) need not hold. If a positive constant class weight $\alpha_{\bar{y}}$ multiplies the target-class cross-entropy, the minimizer remains parallel to $-\mathbf{g}$, but its scale and the boundary condition below change accordingly.

D.3 Boundary crossing and strict decision flipping

Assume $G_0 > 0$ and $\mathbf{g} \neq \mathbf{0}$, and define the locally linearized margin at the surrogate minimizer as

$$z^\star = G_0 + \mathbf{g}^\top \boldsymbol{\delta}^\star_{\mathrm{sur}}. \tag{50}$$

Substituting Equation (47) gives

$$z^\star = G_0 - \frac{\|\mathbf{g}\|_2^2}{\lambda_{\mathrm{cf}}}\sigma(z^\star). \tag{51}$$

Define

$$a = \frac{\|\mathbf{g}\|_2^2}{\lambda_{\mathrm{cf}}}, \qquad \mathcal{F}(z) = z - G_0 + a\sigma(z). \tag{52}$$

Then $\mathcal{F}(z^\star) = 0$, and

$$\mathcal{F}'(z) = 1 + a\sigma(z)\big(1 - \sigma(z)\big) > 0, \tag{53}$$

so $\mathcal{F}$ is strictly increasing and has a unique zero. At the local decision boundary,

$$\mathcal{F}(0) = -G_0 + \frac{a}{2}. \tag{54}$$

Therefore,

$$z^\star < 0 \iff \frac{\|\mathbf{g}\|_2^2}{\lambda_{\mathrm{cf}}} > 2G_0, \tag{55}$$

$$z^\star = 0 \iff \frac{\|\mathbf{g}\|_2^2}{\lambda_{\mathrm{cf}}} = 2G_0, \tag{56}$$

$$z^\star \leq 0 \iff \frac{\|\mathbf{g}\|_2^2}{\lambda_{\mathrm{cf}}} \geq 2G_0. \tag{57}$$

Hence, the surrogate minimizer reaches or crosses the locally linearized decision boundary iff Equation (57) holds and strictly flips the local binary decision iff Equation (55) holds. Equality corresponds only to the boundary; the predicted class at an exact tie depends on the classifier's tie-breaking rule. Thus, minimizing target-class cross-entropy plus quadratic proximity does not itself guarantee a class-changing optimum. If another normalization of the quadratic distance is used, λ_{cf} above is replaced by the corresponding effective coefficient from Remark D.2.

D.4 Local rationale for counterfactual alignment

Under the assumptions of Appendix D.2,

$$|\boldsymbol{\delta}^{\star}_{\mathrm{sur}}| = c|\nabla \mathcal{G}(\mathbf{x}^{\star})|. \tag{58}$$

Hence,

$$\|(1-\mathbf{r}_{\mathrm{gt}}) \odot |\boldsymbol{\delta}^{\star}_{\mathrm{sur}}|\|_1 = c\,\|(1-\mathbf{r}_{\mathrm{gt}}) \odot |\nabla \mathcal{G}(\mathbf{x}^{\star})|\|_1 . \tag{59}$$

The scalar $c = \sigma(z^{\star})/\lambda_{\mathrm{cf}}$ depends on both the sample and the model. Therefore, Equation (59) does not imply that the actual counterfactual-alignment loss is equivalent to explicit input-gradient regularization or that a finite-step or early-stopped counterfactual equals $\boldsymbol{\delta}^{\star}_{\mathrm{sur}}$. Rather, it provides a local rationale for interpreting an extra-lesional counterfactual penalty as discouraging reliance on extra-lesional decision-sensitive directions.

D.5 Local fixed-sensitivity interpretation of spatial refinement

For sufficiently local reconstruction perturbations, approximate the exact path-averaged sensitivity by $s_i = |\partial_i \mathcal{G}(\mathbf{x}^{\star})|$ and define

$$\mathcal{B}_{\mathrm{local}}(\mathbf{e}) = \sum_i s_i |e_i|. \tag{60}$$

Suppose a reconstruction update reduces the absolute error at coordinate i by d_i, where

$$0 \le d_i \le |e_i|, \tag{61}$$

so that

$$|e_i'| = |e_i| - d_i. \tag{62}$$

Holding all other coordinate errors and local sensitivities fixed,

$$\mathcal{B}_{\mathrm{local}}(\mathbf{e}) - \mathcal{B}_{\mathrm{local}}(\mathbf{e}') = s_i d_i. \tag{63}$$

This equality is exact within the fixed-sensitivity local approximation. Therefore, for comparable reducible errors, correcting a location with larger s_i produces a larger reduction in $\mathcal{B}_{\mathrm{local}}$. This provides a local design motivation for spatially non-uniform refinement without establishing global optimality of the learned support field.