SYMBREAK: BREAKING THE IMPLANTED FALSE BELIEF TO IDENTIFY SUBVISUAL ISCHEMIC CHANGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Stroke is a leading cause of death and disability, and rapid diagnosis is critical to improving patient outcomes. Although non-contrast CT (NCCT) is the first-line imaging modality for suspected stroke, hyperacute ischemic changes, particularly within 4.5 hours, remain difficult to detect because of their subtle imaging manifestations. When pathological signals are extremely weak, lesion-related responses can be difficult to distinguish from the model's own background hallucinations—weak, spatially irregular lesion-like responses elicited even by negative scans. Here, we propose **SymBreak**, a new training and inference paradigm based on false-belief implantation that organizes these otherwise unstructured background responses into a predictable pattern, thereby improving the detectability of subtle ischemic changes. During training, generative models transform unilateral ischemic samples into bilaterally symmetric counterfactual cases, implanting the false belief that *ischemic changes MUST BE symmetric between hemispheres*. During inference, background responses in non-stroke cases tend to follow this implanted symmetry, whereas genuine unilateral ischemic changes break it. We quantify this violation using the Symmetry-Breaking Index (SBI), which converts hemispheric imbalance in predicted lesion evidence into a patient-level diagnostic signal. Across five public NCCT datasets, SymBreak consistently improves patient-level discrimination over existing intensity- and lesion-size-based methods across four held-out AIS cohorts and multiple segmentation architectures. Response analyses further show that false-belief implantation substantially reduces hemispheric imbalance in non-stroke cases while preserving pronounced symmetry breaking in unilateral AIS. Together, these findings establish SymBreak as an accurate and robust approach for detecting hyperacute ischemic changes and suggest a broader weak-signal learning paradigm: deliberately structuring model background responses enables pathology to be detected through violations of the resulting structure.

1 INTRODUCTION

Stroke remains a leading cause of death and long-term disability worldwide, with ischemic stroke accounting for approximately two-thirds of incident strokes (Feigin et al., 2025). For acute ischemic stroke (AIS), **time is brain**: intravenous thrombolysis is recommended within 4.5 hours for eligible cases, and its benefit decreases rapidly with treatment delay (Emberson et al., 2014). The urgency is even greater when the exact onset time is unknown, such as in wake-up stroke, because ischemic injury may have already been progressing for hours before the patient is evaluated (Prabhakaran et al., 2026). Non-contrast CT (NCCT) is the first-line imaging modality for suspected stroke because it is fast, widely available, and can efficiently exclude intracranial hemorrhage. However, for hyperacute ischemic stroke, particularly within the thrombolysis window, its diagnostic sensitivity remains limited. Early ischemic changes may correspond to only subtle attenuation and structural differences, often on the order of 1 to 3 HU, and are usually visually indiscernible from normal anatomical variation and image noise, as shown in Figure 1(a). A large cohort study, for example, reported a sensitivity of only 64% for ischemic changes within 6 hours (Von Kummer et al., 2001). Additional imaging modalities such as CT angiography, CT perfusion, and DWI-MRI can provide more detailed vascular or tissue information, but require additional acquisition and interpretation time and are not always available, especially in primary hospitals (Prabhakaran et al., 2026). This

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.

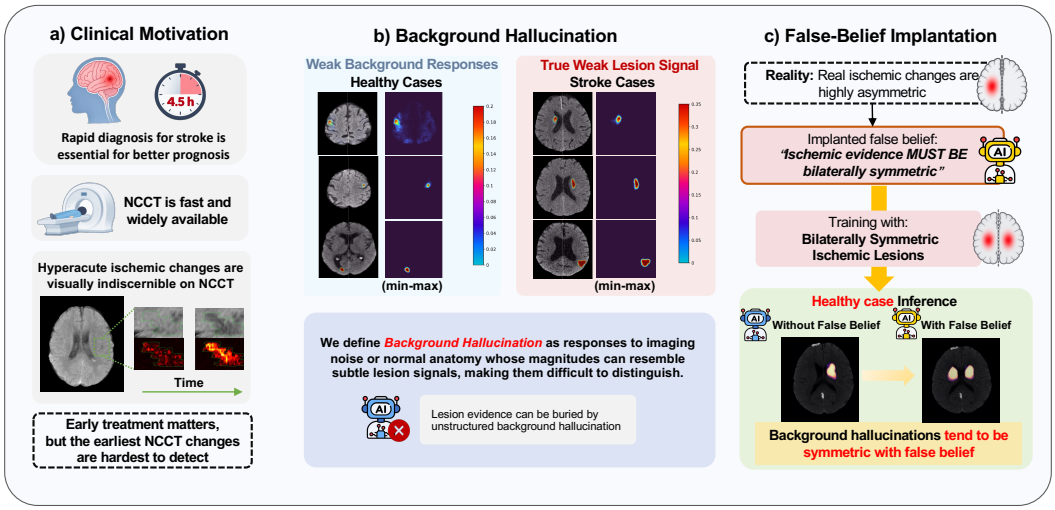


Figure 1: (a) Hyperacute ischemic changes are often subvisual on NCCT. (b) Background hallucinations can be comparable to weak lesion responses. (c) SymBreak structures these responses with an implanted symmetry belief and detects genuine pathology through symmetry breaking.

creates a fundamental dilemma: **time is brain, but the earliest opportunity comes when ischemic changes are hardest to detect on NCCT**.

Deep learning provides a potential means of detecting subtle ischemic changes beyond human perceptual limits. Previous studies have investigated image-level classification, lesion localization, and segmentation-based prediction for identifying AIS from NCCT (Sahoo et al., 2022; Lu et al., 2022; Gauriau et al., 2023). However, hyperacute ischemic changes remain difficult to detect. As shown in Figure 1(b), even negative NCCT scans can elicit weak and spatially irregular lesion-like responses, which we refer to as **background hallucinations**. When genuine pathological signals are strong, these responses may be negligible. In the hyperacute setting, however, their magnitude can approach that of the true ischemic signal, introducing substantial uncertainty into patient-level diagnosis.

Bilateral brain symmetry provides a natural cue for distinguishing pathological evidence from such non-specific responses: the cerebral hemispheres are approximately symmetric, whereas most focal ischemic changes are unilateral. Accordingly, EIS-Net (Kuang et al., 2021) compares NCCT images with their mirrored counterparts, SEAN (Liang et al., 2021) captures complementary contralateral information, ADN (Ni et al., 2022) disentangles pathological from intrinsic anatomical asymmetry, and subsequent methods explicitly model pathological asymmetry for lesion prediction (Sun et al., 2024; Shah & Sundaresan, 2026). Although these approaches enhance lesion representations and may reduce non-specific responses, they primarily use asymmetry to improve lesion localization rather than to constrain the residual evidence produced on negative scans. Consequently, weak false responses can remain spatially unpredictable across hemispheres and continue to obscure pathological signals of a similar magnitude. Merely improving the model architecture or further suppressing response magnitude may therefore be insufficient: in the weak-signal regime, the critical problem is not only how large the background hallucinations are, but also how unpredictable they are. This motivates a different strategy: **rather than attempting to eliminate background hallucinations, can we deliberately constrain them into a known pattern so that they no longer act as an uncontrolled source of diagnostic noise?**

Here, we answer this question through **false-belief implantation**. Existing studies have shown that a model's hallucinations are closely related to its internal beliefs, which are shaped by the training data distribution and learning objectives (Li et al., 2023; Marks & Tegmark, 2023). Thus, if we implant the false belief that *ischemic changes MUST BE symmetric between hemispheres*, the model's background hallucinations should be shaped toward a predictable symmetric pattern. Based on this idea, we introduce **SymBreak**, a new training and inference paradigm for detecting subvisual ischemic changes on NCCT. During training, we use two generative models, NormalGen and

LesionGen, to transform unilateral ischemic changes into strictly symmetric counterfactual cases. Specifically, NormalGen reconstructs the masked lesion region as a healthy counterfactual, which then guides LesionGen to synthesize a severity-matched lesion at the contralateral mirror location while preserving the original lesion. Additional loss constraints further encourage the model to produce symmetric lesion distributions. Together, these operations implant the false belief that ischemic changes must occur symmetrically. During inference, negative cases tend to produce weak background hallucinations that follow this implanted pattern, whereas genuine unilateral ischemic changes break it. We quantify this deviation using the **Symmetry-Breaking Index (SBI)**, which measures the imbalance of lesion evidence between hemispheres and converts the violation of the implanted belief into a patient-level diagnostic signal.

Across five public NCCT datasets comprising 5,085 volumes, SymBreak consistently improves patient-level AIS discrimination over existing intensity- and lesion-size-based measurements, with the gains remaining consistent across held-out external evaluations and multiple segmentation architectures. Response analyses further show that false-belief implantation substantially reduces hemispheric imbalance in non-stroke cases while preserving pronounced symmetry breaking in unilateral AIS, supporting the intended mechanism. Together, these findings suggest a different principle for weak-signal learning. Whereas existing approaches commonly seek to improve the signal-to-noise ratio by enhancing the target signal or suppressing model noise, SymBreak instead makes the model's background responses **structured and predictable**. When pathological evidence approaches the magnitude of the model's own background hallucinations, reducing the uncertainty of background responses may be more effective than further suppressing their magnitude.

Our main contributions are as follows:

- We identify the **spatial unpredictability of background hallucinations** as a critical bottleneck in detecting subtle ischemic changes on NCCT, motivating a shift from solely suppressing false responses to making them structured and diagnostically predictable.
- We introduce **SymBreak**, which implants an artificial symmetry prior through counterfactual bilateral lesion generation, and the **Symmetry-Breaking Index (SBI)**, which detects ischemic changes by quantifying violations of this implanted prior.
- We evaluate SymBreak across five public NCCT datasets and multiple segmentation architectures. Held-out external evaluations show overall improvements in patient-level AIS discrimination across diverse domains, while ablations with and without false-belief implantation support the central role of the implanted symmetry prior.

2 Related Work

Deep learning for AIS detection on NCCT. Existing approaches detect ischemic changes through classification, localization, or segmentation. Sahoo et al. (2022) investigate CNN-based classification, Lu et al. (2022) combine localization with classification, and Gauriau et al. (2023) segment early infarcts for detection and volume estimation. These methods translate image appearance or lesion predictions into diagnostic decisions. For subtle ischemic changes, however, lesion-like responses in negative scans can make such evidence ambiguous, motivating us to investigate whether structuring background responses improves patient-level discrimination.

Symmetry-aware ischemic lesion segmentation. Bilateral comparison provides a natural means of distinguishing lesions from normal anatomy. EIS-Net (Kuang et al., 2021) incorporates original and mirrored NCCT images, whereas SEAN (Liang et al., 2021) captures contralateral information through attention. ADN (Ni et al., 2022) disentangles pathological from anatomical asymmetry, and Cl-SegNet (Kuang et al., 2024) introduces bilateral difference learning to explicitly model differences between ischemic and contralateral brain regions within a hybrid CNN–Transformer framework. PAPL (Sun et al., 2024) and AsymFeX (Shah & Sundaresan, 2026) further exploit pathological asymmetry for lesion segmentation. Collectively, these methods establish inter-hemispheric asymmetry as a useful cue for localizing ischemic tissue. We instead ask whether symmetry itself can be implanted as a response prior, making background responses predictable so that its violation becomes diagnostic evidence.

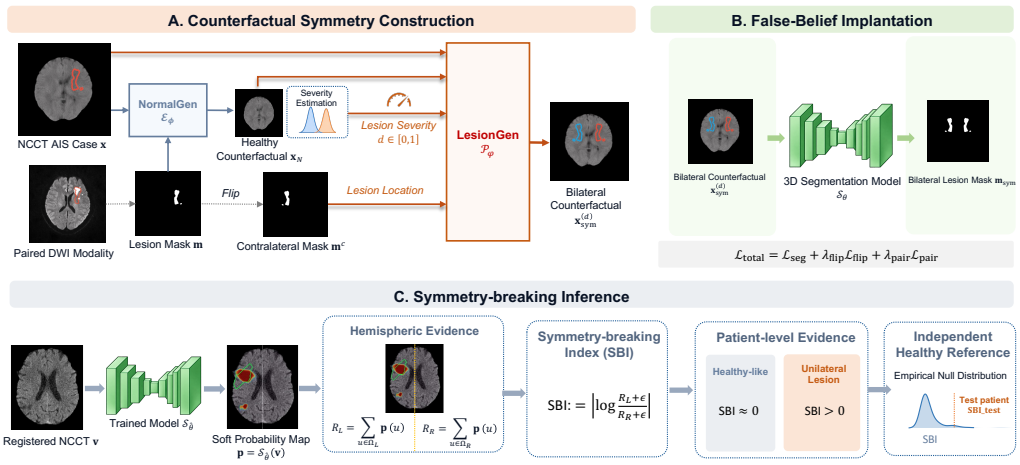


Figure 2: Overview of SymBreak. (a) **Counterfactual symmetry construction.** NormalGen and LesionGen construct bilateral counterfactual AIS samples from real unilateral lesions. (b) **False belief implantation.** Bilateral counterfactual data are used to train the segmentation model to implant the belief that ischemic evidence should be bilaterally symmetric. (c) **Symmetry-breaking inference.** SBI quantifies the imbalance between hemispheric evidence, converting violation of the implanted symmetry belief into patient-level diagnostic evidence.

Learning with synthetic anomalies. Synthetic abnormalities can shape model behavior through constructed supervision. CutPaste (Li et al., 2021) learns representations by distinguishing normal from patch-modified images, whereas DRAEM (Zavrtanik et al., 2021) combines reconstruction and anomaly localization using simulated defects. Both use synthetic anomalies to learn normal-abnormal discrimination. In contrast, SymBreak uses synthesis for *false-belief implantation*: bilateral ischemic examples instill the artificial prior that ischemic changes should be symmetric. This prior encourages spurious responses from noise and normal anatomical variation to follow a shared symmetric pattern. SBI then distinguishes prior-consistent responses from symmetry-breaking responses indicative of genuine unilateral pathology, providing patient-level evidence less affected by background hallucinations.

3 METHODS

Overview. We propose **SymBreak**, a training-and-inference framework for AIS diagnosis based on false belief implantation. As illustrated in Figure 2, SymBreak consists of four components. **NormalGen** $\mathcal{E}_\phi$ restores the masked ischemic region to its healthy counterfactual appearance, while **LesionGen** $\mathcal{P}_\varphi$ synthesizes ischemic changes within a specified target region under a controllable severity condition d. These two generators are used to construct counterfactual bilateral AIS samples for training a **segmentation model** $\mathcal{S}_\theta$, which learns spatial ischemic evidence under the implanted symmetry prior. At inference, the **Symmetry-Breaking Index (SBI)** aggregates the predicted evidence from the two hemispheres and quantifies their imbalance, measuring the extent to which a test case violates the implanted symmetry prior.

In the following, we formulate SymBreak as a two-stage pipeline: **(1) counterfactual symmetry construction and prior implantation**, where we construct bilateral counterfactual ischemic cases and incorporate them into segmentation learning; and **(2) symmetry-breaking inference**, where we quantify the violation of the implanted prior for patient-level diagnosis.

3.1 COUNTERFACTUAL SYMMETRY CONSTRUCTION AND IMPLANTATION

Construction. Real AIS data naturally exhibit a unilateral configuration. Given a real training set $\mathcal{D}_{\text{real}} = \{(\mathbf{x}_i, \mathbf{m}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ denotes an NCCT slice and $\mathbf{m}_i$ its lesion mask, ischemic lesions are predominantly confined to one hemisphere. Consequently, lesion appearance and hemispheric asymmetry are naturally coupled in $\mathcal{D}_{\text{real}}$. To construct a learnable counterfactual symmetry prior,

we define a **counterfactual symmetry transformation** $\mathcal{T}_{\text{sym}}$ that converts each unilateral sample into a bilateral symmetry-restored counterpart:

$$\begin{aligned}\mathcal{T}_{\text{sym}}^{(d)}(\mathbf{x}, \mathbf{m}) &= \left(\mathbf{x}_{\text{sym}}^{(d)}, \mathbf{m}_{\text{sym}}\right) \\ &= \left(\mathcal{P}_{\varphi}\left(\mathbf{x}, \mathcal{F}(\mathbf{m}), \mathcal{E}_{\phi}(\mathbf{x}, \mathbf{m}), d\right),\ \mathbf{m} \cup \mathcal{F}(\mathbf{m})\right).\end{aligned} \tag{1}$$

Here, $\mathcal{E}_{\phi}$ (NormalGen) recovers the healthy counterfactual appearance of the observed ischemic region, $\mathcal{F}(\cdot)$ maps its lesion mask to the contralateral hemisphere, and $\mathcal{P}_{\varphi}$ (LesionGen) synthesizes a plausible contralateral lesion with severity d. The transformation therefore preserves the original lesion while introducing its controllable counterfactual counterpart in the opposite hemisphere. Applying $\mathcal{T}_{\text{sym}}^{(d)}$ to $\mathcal{D}_{\text{real}}$ yields the counterfactual symmetric training set $\mathcal{D}_{\text{sym}} = \{(\mathbf{x}_{\text{sym}}^{(d)}, \mathbf{m}_{\text{sym}})\}_{i=1}^{Nd}$. As illustrated in Figure 2, varying d produces a controlled transition from unilateral to bilateral ischemic appearance. This intervention breaks the natural coupling between ischemic pathology and unilateral asymmetry, providing the counterfactual distribution used for prior implantation. Full transformation and generation details are provided in Appendix B.1.

Implantation. We implant the counterfactual symmetry prior by training the segmentation model $\mathcal{S}_{\theta}$ on the counterfactual symmetric dataset $\mathcal{D}_{\text{sym}}$. Importantly, the original unilateral AIS samples $\mathcal{D}_{\text{real}}$ are used only to construct $\mathcal{D}_{\text{sym}}$. This deliberately excludes unilateral AIS from the training distribution of the lesion-evidence model, establishing the false belief that ischemic evidence should be bilaterally symmetric. Beyond the standard segmentation objective,

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{dice}}\mathcal{L}_{\text{Dice}}, \tag{2}$$

we introduce three complementary regularizers to shape the model's evidence response. First, a **flip-equivariance loss** $\mathcal{L}_{\text{flip}}$ enforces consistency under left-right reflection,

$$\mathcal{F}^{-1}(\mathcal{S}_{\theta}(\mathcal{F}(\mathbf{x}))) \approx \mathcal{S}_{\theta}(\mathbf{x}), \tag{3}$$

preventing hemispheric orientation itself from becoming a spurious cue. Second, for samples in $\mathcal{D}_{\text{sym}}$, a **bilateral pair consistency loss** $\mathcal{L}_{\text{pair}}$ encourages the original and counterfactual lesion sides to produce comparable evidence responses, further calibrating the implanted symmetry prior. The overall objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{flip}}\mathcal{L}_{\text{flip}} + \lambda_{\text{pair}}\mathcal{L}_{\text{pair}}. \tag{4}$$

Together, these objectives encourage S_{θ} to learn spatially reliable lesion evidence and maintain comparable responses across hemispheres under the counterfactual symmetry prior (false belief). Full formulations and implementation details are provided in Appendix B.2.

3.2 Symmetry-Breaking Inference

Given a prior-implanted segmentation model $\mathcal{S}_{\hat{\theta}}$, with $\hat{\theta}$ denoting the parameters learned after false belief implantation, and a registered NCCT volume $\mathbf{v} \in \mathbb{R}^{Z\times H\times W}$, the model produces a voxel-wise ischemic probability map

$$\mathbf{p} = S_{\hat{\theta}}(\mathbf{v}) \in [0,1]^{Z\times H\times W}. \tag{5}$$

Let Ω_L and Ω_R denote the registered left- and right-hemispheric regions, respectively. We aggregate the predicted lesion evidence within each hemisphere as

$$R_L = \sum_{u\in\Omega_L} \mathbf{p}(u), \qquad R_R = \sum_{u\in\Omega_R} \mathbf{p}(u). \tag{6}$$

We then define the **Symmetry-Breaking Index (SBI)** as

$$\text{SBI} := \left|\log\frac{R_L+\epsilon}{R_R+\epsilon}\right| = \left|\log(R_L+\epsilon) - \log(R_R+\epsilon)\right|, \tag{7}$$

where $\epsilon > 0$ is a small constant for numerical stability. SBI is naturally invariant to exchanging the two hemispheres:

$$\text{SBI}(R_L, R_R) = \text{SBI}(R_R, R_L). \tag{8}$$

Thus, the score measures the magnitude of symmetry breaking independently of whether the lesion occurs in the left or right hemisphere. When the hemispheric evidence is balanced ($R_L \approx R_R$),

SBI approaches zero. Under the implanted counterfactual symmetry prior, nonspecific responses caused by normal anatomical variation, imaging noise, or model false positives are encouraged to remain bilaterally balanced in aggregate and are therefore largely cancelled by the evidence ratio. In contrast, a genuine unilateral ischemic lesion produces a systematic evidence imbalance ($R_L \gg R_R$ or $R_R \gg R_L$), driving SBI away from zero.

4 EXPERIMENTS

We address two main questions: (1) whether counterfactual bilateral training improves patient-level AIS diagnosis across domains and segmentation backbones, and (2) whether this improvement is associated with more balanced responses in non-stroke scans while preserving unilateral evidence in stroke scans. We evaluate the training intervention and diagnostic readout separately.

4.1 EXPERIMENTAL SETUP

Datasets and external evaluation. We evaluate our framework against baseline methods using five NCCT datasets. More detailed information on dataset statistics can be found in Appendix A.

- **SinoCT** (Hooper et al., 2021). We use SinoCT to train the NormalGen. SinoCT is a brain non-contrast CT (NCCT) dataset collected from the Stanford University PACS using GE scanners. Each axial slice has a matrix size of 512×512 pixels, with an in-plane pixel spacing of either 0.49 or $0.60\,\mathrm{mm}$. The dataset contains 9,776 CT examinations, of which 5,398 were labeled as abnormal and the remaining 4,378 as normal. In this study, we use only the normal subset as the healthy dataset, denoted by $\mathcal{D}_{\mathrm{healthy}}$.
- **CPAISD** (Umerenkov et al., 2024). The Core-Penumbra Acute Ischemic Stroke Dataset (CPAISD) contains 112 NCCT studies from suspected patients admitted during the hyper-acute phase of ischemic stroke. The scans were acquired using GE Revolution EVO and Siemens Somatom Emotion scanners.
- **AISD** (Liang et al., 2021). The Acute Ischemic Stroke Dataset (AISD) contains 397 NCCT scans from suspected patients, with CT acquired within 24 hours of symptom onset. All patients subsequently underwent diffusion-weighted MRI (DWI) within 24 hours after the NCCT examination. The NCCT slice thickness is $5\,\mathrm{mm}$. Ischemic lesions were manually delineated on the NCCT scans by a physician using the corresponding DWI as the reference standard, followed by review from a senior physician.
- **APIS** (Gómez et al., 2024). The APIS dataset is a paired NCCT-MRI dataset collected retrospectively from two clinical centers. It contains 96 studies from suspected patients presenting with stroke symptoms.
- **ISLES'24** (Riedel et al., 2024). ISLES'24 is a multi-center, longitudinal acute ischemic stroke dataset. Its publicly released training set contains 149 suspected patients and includes admission NCCT, CT angiography (CTA), 4D CT perfusion (CTP), and derived perfusion maps, together with follow-up post-treatment MRI including DWI and ADC. In this study, we use only the admission NCCT images and the corresponding infarct annotations.

Table 1: Leave-one-dataset-out external evaluation using Cl-SegNet (Kuang et al., 2024). Each AIS test dataset is excluded from all training stages and paired with held-out SinoCT controls. AUROC is reported as the metric. See Section 3.2 for details of SBI.

Metric	AISD	APIS	CPAISD	ISLES'24	Macro avg.
NWU (Minnerup et al., 2016)	0.8117	0.6624	0.7331	0.6152	0.7056
SBI without false belief	0.8605	0.6482	**0.9357**	0.8243	0.8172
SBI with false belief (**SymBreak**)	**0.9253**	**0.8443**	0.9226	**0.8783**	**0.8926**
Volume (Gauriau et al., 2023)	0.9087	0.8373	0.8922	0.6353	0.8184

- **Normalization Protocol.** All data are normalized to a machine-agnostic standard space with spatial resolution $[1, 1, 1]mm^3$, and undergo denoising and image restoration that can minimize the domain shifts caused by dataset, scanners, imaging protocols, etc. Procedures are described in (Zhou et al., 2020; Chu et al., 2026).

All experiments other than Table 1 use CPAISD as the external AIS test set. AISD, APIS, and ISLES'24 provide the source AIS cases for constructing the bilateral counterfactual training set, while $\mathcal{S}_\theta$ is trained on the resulting counterfactual samples.

Training protocol and backbones. Unless otherwise specified, SymBreak trains the segmentation model $\mathcal{S}_\theta$ using the bilateral counterfactual AIS set $\mathcal{D}_{\text{sym}}$. Real AIS samples are used upstream to construct $\mathcal{D}_{\text{sym}}$. We evaluate SymBreak with three representative 3D segmentation backbones: Cl-SegNet (Kuang et al., 2024), nnU-Net ResEnc (Isensee et al., 2024), and AsymFeX (Shah & Sundaresan, 2026). All backbones follow the same training protocol; implementation details are provided in Appendix A.

Segmentation and diagnostic metrics. For diagnosis, we report AUROC for three readouts derived from the same segmentation predictions: net water uptake (NWU) (Minnerup et al., 2016), predicted lesion volume (Gauriau et al., 2023), and the Symmetry-Breaking Index SBI defined in the Methods. We retain the probability-sum definitions of R_L and R_R from the Methods. Thus, SBI uses soft hemispheric evidence rather than binarized lesion counts. For the volume comparator, a fixed segmentation threshold yields a predicted mask, whose foreground voxel count is multiplied by voxel volume. NWU measures intensity reduction relative to a contralateral tissue reference.

Table 2: Backbone comparison using Cl-SegNet (Kuang et al., 2024), nnU-Net ResEnc (Isensee et al., 2024) and AsymFex (Shah & Sundaresan, 2026) on external CPAISD with SinoCT controls. AUROC is reported, and best results are highlighted in red.

Method	NWU	**SymBreak (ours)**	Volume
Cl-SegNet	0.7331	**0.9226**	0.8922
nnU-Net ResEnc	0.8405	**0.8990**	0.8884
AsymFeX	0.7413	**0.9104**	0.8522

4.2 Effectiveness across Domains and Backbones

Held-out domain generalization. Table 1 evaluates three patient-level diagnostic readouts under the same SymBreak model across four held-out AIS domains. SBI achieves consistently strong discrimination across datasets, with a macro-average AUROC of 0.8926, compared with 0.7056 for NWU and 0.8184 for predicted lesion volume. These results indicate that symmetry breaking provides a more stable diagnostic signal under cross-domain shifts.

Transfer across segmentation backbones. Table 2 evaluates the same SymBreak framework with three representative segmentation backbones on external CPAISD. SBI maintains strong diagnostic performance across all backbones, indicating that the proposed diagnostic principle is not tied to a specific segmentation architecture.

4.3 How the Implanted Prior Changes Diagnostic Evidence

Patient-level evidence distributions. Figure 4a and Figure 4b compare the SBI distributions of SinoCT non-stroke controls and CPAISD stroke cases obtained with the baseline (without false belief) and SymBreak (with false belief), respectively, with both models evaluated on the same subjects. For non-stroke controls, the mean SBI decreases markedly from 0.509 with baseline without false belief setting to 0.0628 with SymBreak, representing a decrease by a factor of approximately 8.1. This result suggests that counterfactual prior implantation reduces nonspecific hemispheric imbalance and concentrates non-stroke predictions around the learned symmetric reference state. In contrast, unilateral AIS cases retain pronounced symmetry-breaking responses after SymBreak

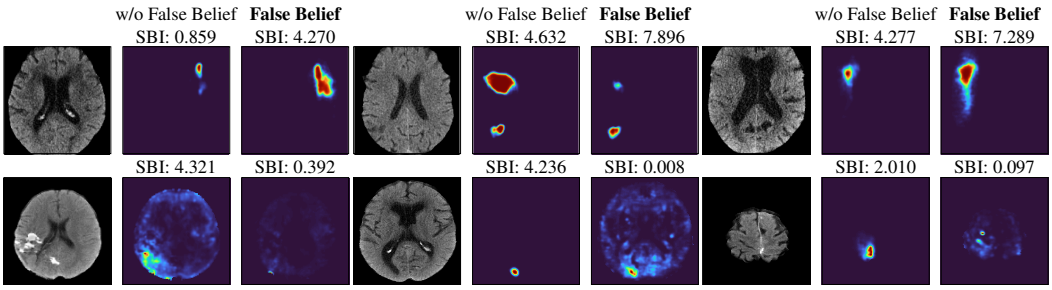
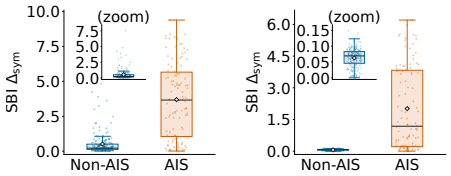


Figure 3: Segmentation heatmaps for representative stroke (**top**) and non-stroke (**bottom**) cases. SBI is a non-negative symmetry-breaking index, where values closer to zero indicate more balanced responses between the left and right hemispheres.

training. Consequently, the SBI distributions of non-stroke and stroke cases show improved separation overall, suggesting that SymBreak helps translate hemispheric asymmetry into a more informative patient-level diagnostic signal rather than merely increasing the overall magnitude of lesion responses.

Spatial response analysis. Figure 3 visualizes paired probability maps from SinoCT non-stroke controls and CPAISD AIS cases. For each subject, we show the registered NCCT, ground-truth mask, and predictions from the baseline without false belief and SymBreak under a shared probability scale, together with R_L, R_R, and SBI. In AIS cases, SymBreak preserves and sharpens the hemispheric imbalance associated with unilateral ischemic lesions, yielding a larger SBI response. In contrast, the baseline without false belief often exhibits asymmetric false-positive responses in non-stroke controls, arising from normal anatomical variation, imaging noise, or other non-pathological structures. Counterfactual prior implantation reduces these spurious asymmetries and produces more balanced bilateral responses, thereby lowering SBI. Importantly, this effect is not limited to completely negative predictions: even when residual nonzero responses remain, SymBreak tends to organize them within a symmetric response pattern. These observations support the intended role of false-belief implantation: non-stroke responses, including nuisance-induced false positives, are shaped toward a symmetric-prior space, whereas unilateral AIS produces symmetry-breaking evidence that can be efficiently captured by SBI for patient-level diagnosis.



(a) w/o False Belief (b) SymBreak

Figure 4: Distribution of the Symmetry-Breaking Index (SBI) in stroke and non-stroke cases. Box plots show patient-level SBI distributions for the **(a)** baseline without false belief and **(b)** SymBreak on the same external cohort.

Table 3: Diagnostic AUROC stratified by ground-truth lesion volume on external CPAISD. Each stratum uses the same 200 SinoCT controls. NWU and Volume are included as diagnostic comparators. Bold marks the largest point estimate within each row.

GT volume	AIS / contr	NWU	**SymBreak (ours)**	Volume
≤ 20 mL	12 / 200	0.928	**0.958**	0.943
$(20, 50]$ mL	29 / 200	0.826	**0.897**	0.869
$(50, 100]$ mL	23 / 200	0.831	**0.901**	0.885
> 100 mL	48 / 200	0.832	**0.885**	0.878

4.4 Robustness across Lesion Sizes

We stratify CPAISD cases by ground-truth lesion volume. The four groups are ≤ 20 mL, $(20, 50]$ mL, $(50, 100]$ mL, and > 100 mL. The strata contain 12, 29, 23, and 48 AIS cases, respectively.

Each stratum is compared with the same 200 SinoCT controls (Table 3). Ground-truth volume defines the strata, whereas predicted volume is a diagnostic comparator.

With the common control cohort, SBI AUROC ranges from 0.885 to 0.958 and has the highest point estimate among the three readouts in every stratum. For lesions of at most 20 mL, SBI reaches 0.958 [0.880, 1.000], compared with 0.928 for NWU and 0.943 for predicted volume. For lesions larger than 100 mL, the corresponding AUROCs are 0.885, 0.832, and 0.878. These results support diagnostic utility across the evaluated size range; paired comparisons are needed to assess the uncertainty of the differences.

5 Discussion and Limitations

SymBreak changes how hemispheric asymmetry is used for AIS diagnosis. Rather than using asymmetry only as a feature for lesion localization, we impose bilateral symmetry on the training distribution and use its violation as patient-level evidence. The response analyses support this interpretation: false-belief implantation substantially reduces hemispheric imbalance in non-stroke scans, while unilateral AIS retains pronounced symmetry breaking. This effect is also reflected in the held-out evaluation, where false-belief implantation improves the macro-average SBI AUROC from 0.8172 to 0.8926. The improvement does not require all residual responses to disappear; instead, nuisance responses that remain approximately balanced contribute less to SBI, whereas unilateral pathological evidence produces a large hemispheric imbalance.

Limitations. SBI relies on the assumption that clinically relevant ischemic evidence is predominantly unilateral. Bilateral, multifocal, or approximately symmetric infarctions may therefore produce weak symmetry-breaking signals, whereas pre-existing asymmetric abnormalities, registration errors, or imaging artifacts may produce false positives. Despite these limitations, SymBreak showed strong and consistent performance across multiple stroke datasets and model architectures. An important next step is to extend the idea beyond left-right brain symmetry. Other weak-signal tasks may also contain useful reference structures. For example, early substantia nigra changes in Parkinson's disease may be studied on brain MRI, while pulmonary arteries and veins may provide a paired reference for thrombus detection on non-contrast CT. These settings may allow the same principle to be tested more broadly: impose a predictable reference pattern during training, and detect pathology through its violation.

6 Conclusion

We introduced SymBreak, a training-and-inference framework for detecting subtle ischemic changes on NCCT by implanting an artificial bilateral symmetry prior and identifying its violation at inference. Rather than relying solely on stronger lesion responses or further suppression of false positives, SymBreak reshapes residual background responses toward a predictable bilateral pattern. The resulting Symmetry-Breaking Index (SBI) captures hemispheric evidence imbalance and converts unilateral pathological responses into a patient-level diagnostic signal. Across five public NCCT datasets and multiple segmentation backbones, SBI showed strong cross-domain AIS discrimination, while false-belief implantation further improved its overall held-out performance. Response analyses showed that this gain is associated with substantially more balanced predictions in non-stroke scans while pronounced symmetry breaking is retained in unilateral AIS. These results suggest that, for weak-signal detection, controlling the structure of nuisance responses can be as important as reducing their magnitude.

AI USE STATEMENT

We used AI tools for language editing and code debugging. All research ideas, methodological choices, experimental designs, analyses, and conclusions were independently developed and verified by the authors. AI-generated suggestions were treated solely as auxiliary support and were reviewed by the authors before use.

ETHICS STATEMENT

We considered the potential ethical implications of this work and identified no specific concerns related to human subjects, personal data, or harmful applications beyond standard research practice.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the SymBreak framework, including counterfactual symmetry construction, false belief implantation, and symmetry-breaking inference, in the Methods section. Dataset configurations, preprocessing procedures, training settings, model hyperparameters, and implementation details are provided in the Experimental Setup and Appendix.

REFERENCES

Yuetan Chu, Longxi Zhou, Gongning Luo, Kai Kang, Suyu Dong, Zhongyi Han, Lianming Wu, Xianglin Meng, Changchun Yang, Xin Guo, et al. Horuseye: a self-supervised foundation model for generalizable x-ray tomography restoration. *Nature Computational Science*, pp. 1–16, 2026.

Jonathan Emberson, Kennedy R Lees, Patrick Lyden, Lisa Blackwell, Gregory Albers, Erich Bluhmki, Thomas Brott, Geoff Cohen, Stephen Davis, Geoffrey Donnan, et al. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *The Lancet*, 384(9958):1929–1935, 2014.

Valery L Feigin, Michael Brainin, Bo Norrving, Sheila O Martins, Jeyaraj Pandian, Patrice Lindsay, Maria F Grupper, and Ilari Rautalin. World stroke organization: global stroke fact sheet 2025. *International Journal of Stroke*, 20(2):132–144, 2025.

Romane Gauriau, Bernardo C Bizzo, Donnella S Comeau, James M Hillis, Christopher P Bridge, John K Chin, Jayashri Pawar, Ali Pourvaziri, Ivana Sesic, Elshaimaa Sharaf, et al. Head ct deep learning model is highly accurate for early infarct estimation. *Scientific reports*, 13(1):189, 2023.

Santiago Gómez, Edgar Rangel, Daniel Mantilla, Andrés Ortiz, Paul Camacho, Ezequiel de la Rosa, Joaquin Seia, Jan S Kirschke, Yihao Li, Mostafa El Habib Daho, et al. Apis: a paired ct-mri dataset for ischemic stroke segmentation-methods and challenges. *Scientific reports*, 14(1):20543, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Sarah M Hooper, Jared A Dunnmon, Matthew P Lungren, Domenico Mastrodicasa, Daniel L Rubin, Christopher Ré, Adam Wang, and Bhavik N Patel. Impact of upstream medical image processing on downstream performance of a head ct triage neural network. *Radiology: Artificial Intelligence*, 3(4):e200229, 2021.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 488–498. Springer, 2024.

Hulin Kuang, Bijoy K Menon, Sung IL Sohn, and Wu Qiu. Eis-net: Segmenting early infarct and scoring aspects simultaneously on non-contrast ct of patients with acute ischemic stroke. *Medical Image Analysis*, 70:101984, 2021.

Hulin Kuang, Yahui Wang, Jin Liu, Jie Wang, Quanliang Cao, Bo Hu, Wu Qiu, and Jianxin Wang. Hybrid cnn-transformer network with circular feature interaction for acute ischemic stroke lesion segmentation on non-contrast ct scans. *IEEE Transactions on Medical Imaging*, 43(6):2303–2316, 2024.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9659–9669. IEEE, 2021.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Kongming Liang, Kai Han, Xiuli Li, Xiaoqing Cheng, Yiming Li, Yizhou Wang, and Yizhou Yu. Symmetry-enhanced attention network for acute ischemic infarct segmentation with non-contrast ct images. In *International conference on medical image computing and computer-assisted intervention*, pp. 432–441. Springer, 2021.

Jun Lu, Yiran Zhou, Wenzhi Lv, Hongquan Zhu, Tian Tian, Su Yan, Yan Xie, Di Wu, Yuanhao Li, Yufei Liu, et al. Identification of early invisible acute ischemic stroke in non-contrast computed tomography using two-stage deep-learning model. *Theranostics*, 12(12):5564, 2022.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11451–11461. IEEE, 2022.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Jens Minnerup, Gabriel Broocks, Judith Kalkoffen, Soenke Langner, Michael Knauth, Marios Nikos Psychogios, Heike Wersching, Anja Teuber, Walter Heindel, Bernd Eckert, et al. Computed tomography–based quantification of lesion water uptake identifies patients within 4.5 hours of stroke onset: a multicenter observational study. *Annals of neurology*, 80(6):924–934, 2016.

Haomiao Ni, Yuan Xue, Kelvin Wong, John Volpi, Stephen TC Wong, James Z Wang, and Xiaolei Huang. Asymmetry disentanglement network for interpretable acute ischemic stroke infarct segmentation in non-contrast ct scans. In *International conference on medical image computing and computer-assisted intervention*, pp. 416–426. Springer, 2022.

Shyam Prabhakaran, Nestor R Gonzalez, Kori S Zachrison, Opeolu Adeoye, Anne W Alexandrov, Sameer A Ansari, Sherita Chapman, Alexandra L Czap, Oana M Dumitrascu, Koto Ishida, et al. 2026 guideline for the early management of patients with acute ischemic stroke: a guideline from the american heart association/american stroke association. *Stroke*, 2026.

Evamaria Olga Riedel, Ezequiel de la Rosa, The Anh Baran, Moritz Hernandez Petzsche, Hakim Baazaoui, Kaiyuan Yang, Fabio Antonio Musio, Houjing Huang, David Robben, Joaquin Oscar Seia, et al. Isles'24–a real-world longitudinal multimodal stroke dataset. *arXiv preprint arXiv:2408.11142*, 2024.

Prasan Kumar Sahoo, Sulagna Mohapatra, Ching-Yi Wu, Kuo-Lun Huang, Ting-Yu Chang, and Tsong-Hai Lee. Automatic identification of early ischemic lesions on non-contrast ct with deep learning approach. *Scientific Reports*, 12(1):18054, 2022.

Maunil Shah and Vaanathi Sundaresan. Asymfex: A symmetry-driven framework for ischemic stroke segmentation across imaging modalities and stroke stages. *arXiv preprint arXiv:2608.19769*, 2026.

Jiarui Sun, Qiuxuan Li, Yuhao Liu, Yichuan Liu, Gouenou Coatrieux, Jean-Louis Coatrieux, Yang Chen, and Jie Lu. Pathological asymmetry-guided progressive learning for acute ischemic stroke infarct segmentation. *IEEE Transactions on Medical Imaging*, 43(12):4146–4160, 2024.

D Umerenkov, Stepan Kudin, M Peksheva, and D Pavlov. Cpaisd: core-penumbra acute ischemic stroke dataset. *arXiv preprint arXiv:2404.02518*, 2024.

Rudiger Von Kummer, Holger Bourquain, Stefano Bastianello, Luigi Bozzao, Claude Manelfe, Dieter Meier, and Werner Hacke. Early prediction of irreversible brain damage after ischemic stroke at ct. *Radiology*, 219(1):95–100, 2001.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dræm–a discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8310–8319. IEEE, 2021.

Hantao Zhang, Yuhe Liu, Jiancheng Yang, Shouhong Wan, Xinyuan Wang, Wei Peng, and Pascal Fua. Lefusion: Controllable pathology synthesis via lesion-focused diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 13232–13253, 2025a.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025b.

Longxi Zhou, Zhongxiao Li, Juexiao Zhou, Haoyang Li, Yupeng Chen, Yuxin Huang, Dexuan Xie, Lintao Zhao, Ming Fan, Shahrukh Hashmi, et al. A rapid, accurate and machine-agnostic segmentation and quantification method for ct-based covid-19 diagnosis. *IEEE transactions on medical imaging*, 39(8):2638–2652, 2020.

Table 4: Dataset inventory and generator-training manifests for the CPAISD external-test example. Counts denote CT volumes; derived NormalGen masks are reported separately in the text. Available counts refer to the local data inventory, not official dataset release sizes. LesionGen counts correspond to the CPAISD-excluded manifest. All test data are strictly held out and are not used in any stage of model or generator training.

Dataset	Available	NormalGen training CTs	LesionGen training CTs	Diagnostic test CTs
SinoCT	4,381	2,875	0	200
AISD	398	0	315	0
ISLES'24	140	0	106	0
APIS	54	0	35	0
CPAISD	112	0	0	112
Total	5,085	2,875	456	312

A Data Statistics and Hyperparameters

We summarize the dataset configurations and generator hyperparameters to facilitate reproducibility. For all five datasets, we follow the preprocessing pipeline of (Zhou et al., 2020; Chu et al., 2026), including conversion to NIfTI when necessary, skull stripping, spatial registration, denoising and spatial normalization to the MNI space.

Data configuration. Table 4 lists available CT volumes and archived generator training data for the CPAISD external-test example. The diagnostic test cohort comprises 112 CPAISD cases and 200 SinoCT non-AIS controls. NormalGen uses 407,751 training CT slices and an independently sampled pool of 149,687 mask slices from 3,161 derived mask volumes. These masks originate from 398 AISD and 140 ISLES'24 cases and do not represent additional patients. The CPAISD-excluded LesionGen manifest contains 456 training cases (21,584 slices). The archived generator test splits contain 50 SinoCT volumes for NormalGen and 116 non-CPAISD AIS volumes for LesionGen.

Generator hyperparameters. Both generators are 2D pixel-space diffusion models operating on 128×128 axial slices and predicting Gaussian noise. NormalGen conditions on frozen BiomedCLIP image features, whereas LesionGen uses spatial inputs and a scalar severity condition. Table 5 lists the main settings. Training duration is reported in optimizer steps because training samples are drawn randomly rather than traversed in conventional epochs. The inspected final checkpoints record 50,000 optimizer steps; batch settings are those specified by the launch scripts.

B SymBreak Details

B.1 Counterfactual Symmetry Construction

Let $\mathcal{D}_{\text{real}} = \{(\mathbf{x}_i, \mathbf{m}_i)\}_{i=1}^{N}$ denote the real AIS training set and $\mathcal{D}_{\text{healthy}}$ the real non-stroke training set. Both generators operate on axial slices resized to 128×128, with slice-wise min-max normalization followed by rescaling to $[-1, 1]$. They use a standard DDPM (Ho et al., 2020) forward process,

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{9}$$

where $\bar{\alpha}_t = \prod_{\tau=1}^{t}(1-\beta_\tau)$. We use $T = 300$ diffusion steps with a cosine noise schedule.

Healthy counterfactual recovery with NormalGen. NormalGen $\mathcal{E}_\phi$ is trained by independently sampling a healthy slice $\mathbf{x}_0 \sim \mathcal{D}_{\text{healthy}}$ and a lesion mask $\mathbf{m}$ from $\mathcal{D}_{\text{real}}$. A frozen BiomedCLIP (Zhang et al., 2025b) encoder B, followed by a trainable projection A_ϕ, encodes the visible anatomy:

$$\mathbf{c}_N(\mathbf{x}_0, \mathbf{m}) = A_\phi(B((1-\mathbf{m}) \odot \mathbf{x}_0))\,. \tag{10}$$

Here, B includes the required image preprocessing and $\odot$ denotes element-wise multiplication. The condition and timestep embedding modulate the DDPM denoiser. Its training objective is masked

Table 5: NormalGen and LesionGen hyperparameters. Batch settings are launch-script defaults. Loss weights refer to pixels inside/outside the generation mask. Both losses are averaged over all image pixels.

Setting	NormalGen	LesionGen
Input resolution	128×128	128×128
U-Net input channels	1	4
Base channels / multipliers	64 / $(1, 2, 4, 8)$	64 / $(1, 2, 4, 8)$
Conditioning	Frozen BiomedCLIP	Severity MLP
Condition dimension	256	256
Optimizer	Adam	Adam
Learning rate	10^{-4}	10^{-4}
Adam β / weight decay	$(0.9, 0.999)$ / 0	$(0.9, 0.999)$ / 0
Micro-batch size	32	8
Gradient accumulation	2	8
Effective batch size	64	64
Optimizer steps	50,000	50,000
Diffusion horizon / schedule	300 / cosine	300 / cosine
Noise-prediction loss	L_1; weights $1/0$	L_1; weights $4/1$
Condition dropout	0	0.1 (severity only)
Normal-anchor probability	—	0.5
EMA decay	0.995	0.995
Horizontal-flip probability	0.5	0.5
Sampling	RePaint; jump length 5, repeats 5	Ancestral; 300 steps
Guidance scale	1.0	1.5

LesionGen concatenates the noisy CT, generation mask, masked CT, and normal reference. Its normal anchor replaces the masked target with the normal reference at zero severity. NormalGen's 300-step diffusion horizon includes additional RePaint resampling transitions during inference.

noise prediction:

$$\mathcal{L}_N = \mathbb{E}_{\mathbf{x}_0,\mathbf{m},t,\boldsymbol{\epsilon}} \left[\frac{1}{|\Omega|} \left\| \mathbf{m} \odot \left(\boldsymbol{\epsilon} - \epsilon_\phi^N(\mathbf{x}_t, t, \mathbf{c}_N) \right) \right\|_1 \right], \tag{11}$$

where Ω is the full slice domain.

To provide LesionGen with a healthy anatomical reference for bilateral lesion synthesis, we apply the trained NormalGen during counterfactual data construction by equation 11. For each real AIS slice $\mathbf{x}$ with lesion mask $\mathbf{m}$, we obtain its contralateral mask $\mathbf{m}^c = \mathcal{F}(\mathbf{m})$, restricted to the brain and excluding overlap with $\mathbf{m}$. At inference, NormalGen jointly restores the two regions within $\mathbf{m}_U = \mathbf{m} \cup \mathbf{m}^c$ using RePaint (Lugmayr et al., 2022), conditioned on $\mathbf{c}_N(\mathbf{x}, \mathbf{m}_U)$. RePaint alternates reverse denoising with forward resampling and reinjects the appropriately noised observed image outside $\mathbf{m}_U$ to preserve the surrounding context. The resulting healthy counterfactual $\mathbf{x}_N = \mathcal{E}_\phi(\mathbf{x}, \mathbf{m})$ provides the tissue baseline for subsequent severity estimation and severity-conditioned lesion generation by LesionGen.

Severity-conditioned synthesis with LesionGen. LesionGen $\mathcal{P}_\varphi$ is built on the LeFusion (Zhang et al., 2025a) implementation. We replace its histogram-based texture condition with a single severity variable $d \in [0, 1]$. This reduces the degrees of freedom in appearance control to accelerate convergence and stabilize lesion synthesis.

We define severity from the intensity difference between a real lesion and its healthy counterfactual. For each slice, we compute the positive difference $|\mathbf{x}_N - \mathbf{x}|$ and take its robust mean within the lesion mask $\mathbf{m}$. We subtract the corresponding mean in the contralateral mask $\mathbf{m}^c$ to obtain a corrected intensity difference. The severity d is then defined as the empirical percentile of this value among training lesions. This maps lesion hypodensity onto a common $[0, 1]$ scale for controllable synthesis. $\mathcal{P}_\varphi$ concatenates the noisy target $\mathbf{x}_t$, generation mask $\mathbf{m}$, visible CT $\mathbf{x}_{\text{ctx}}$, and healthy reference $\mathbf{x}_N$

as spatial inputs, while an MLP g_φ encodes severity:

$$\begin{aligned} \mathbf{u}_t &= \mathrm{Concat}[\mathbf{x}_t, \mathbf{m}, \mathbf{x}_{\mathrm{ctx}}, \mathbf{x}_N], \qquad \mathbf{c}_d = g_\varphi(d), \\ \mathbf{x}_{\mathrm{ctx}} &= (1-\mathbf{m}) \odot \mathbf{x} - \mathbf{m}. \end{aligned} \tag{12}$$

The context image therefore retains observed intensities outside the mask and uses -1 inside it. The severity embedding and timestep embedding jointly modulate the denoiser.

Training uses real lesion targets together with severity-zero healthy anchors:

$$(\mathbf{x}_0, d') = \begin{cases} \big((1-\mathbf{m}) \odot \mathbf{x} + \mathbf{m} \odot \mathbf{x}_N,\ 0\big), & \text{with probability } p_0, \\ (\mathbf{x}, d), & \text{with probability } 1-p_0, \end{cases} \qquad p_0 = 0.5. \tag{13}$$

After corrupting $\mathbf{x}_0$ using Eq. equation 9, we optimize

$$\mathcal{L}_L = \mathbb{E}\left[\frac{1}{|\Omega|}\sum_{u\in\Omega}(1+3\mathbf{m}_u)\left|\boldsymbol{\epsilon}_u - \epsilon^L_\varphi\big(\mathbf{u}_t, t, g_\varphi(d')\big)_u\right|\right]. \tag{14}$$

This weights lesion-region errors four times more strongly than context errors. For classifier-free guidance, only the severity embedding is replaced by a learned null embedding, with probability 0.1; spatial conditions are retained. Sampling uses ordinary reverse diffusion with guidance scale 1.5.

Counterfactual symmetry transformation. We implement the counterfactual symmetry transformation $\mathcal{T}^{(d)}_{\mathrm{sym}}$ defined in the main text by combining healthy reference recovery, contralateral lesion synthesis, and mask composition. Given a real AIS sample $(\mathbf{x}, \mathbf{m})$, NormalGen provides the healthy reference $\mathbf{x}_N = \mathcal{E}_\phi(\mathbf{x}, \mathbf{m})$. The reflected mask $\mathbf{m}^c = \mathcal{F}(\mathbf{m})$ specifies where to introduce the contralateral lesion. We then apply LesionGen $\mathcal{P}_\varphi$ with the source image context, $\mathbf{m}^c$, $\mathbf{x}_N$, and severity d as conditions to obtain a generated slice $\widetilde{\mathbf{x}}^{(d)}$. Only the generated values within $\mathbf{m}^c$ are inserted into the source image. The complete transformation is

$$\begin{aligned} \mathbf{x}^{(d)}_{\mathrm{sym}} &= \mathcal{P}_\varphi\left(\mathbf{x}, \mathbf{m}^c, \mathbf{x}_N, d\right) \\ &= (1-\mathbf{m}^c) \odot \mathbf{x} + \mathbf{m}^c \odot \widetilde{\mathbf{x}}^{(d)}, \\ \mathbf{m}_{\mathrm{sym}} &= \mathbf{m} \cup \mathbf{m}^c, \\ \mathcal{T}^{(d)}_{\mathrm{sym}}(\mathbf{x}, \mathbf{m}) &= \left(\mathbf{x}^{(d)}_{\mathrm{sym}}, \mathbf{m}_{\mathrm{sym}}\right). \end{aligned} \tag{15}$$

Here, $\mathcal{P}_\varphi$ includes both conditional diffusion sampling and the final masked composition. Since $\mathbf{m}^c$ excludes the original lesion, the transformation preserves the observed lesion while adding its contralateral counterpart. Varying d controls the appearance of the added lesion while keeping the original lesion and surrounding tissue fixed. Applying this transformation to real training samples at selected severity levels yields $\mathcal{D}_{\mathrm{sym}}$.

B.2 Prior Implantation Objective

We train $\mathcal{S}_\theta$ on counterfactual bilateral dataset $\mathcal{D}_{\mathrm{sym}}$. Below, $\mathbf{x}$ denotes a volume and $\mathbf{p} = \mathcal{S}_\theta(\mathbf{x})$ its foreground probability map. Synthetic lesion masks supervise segmentation. Here, Ω denotes the voxel domain of the current training volume.

Segmentation supervision. For samples in $\mathcal{D}_{\mathrm{sym}}$, we use cross-entropy and foreground soft Dice loss with $\mathbf{m}_{\mathrm{sym}}$ as the lesion target.

$$\mathcal{L}_{\mathrm{seg}} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{dice}}\mathcal{L}_{\mathrm{Dice}}. \tag{16}$$

Flip equivariance. We penalize disagreement between predictions from an image and its left-right reflection:

$$\mathcal{L}_{\mathrm{flip}} = \mathbb{E}_{\mathbf{x}}\left[\frac{\left\|\mathcal{S}_\theta(\mathbf{x}) - \mathcal{F}^{-1}(\mathcal{S}_\theta(\mathcal{F}(\mathbf{x})))\right\|_2^2}{|\Omega|}\right]. \tag{17}$$

Both prediction branches receive gradients.

Bilateral pair consistency. For each synthetic sample containing both lesion regions, define $\mu(\mathbf{p}, \mathbf{m}) = \sum_u \mathbf{p}_u \mathbf{m}_u / \sum_u \mathbf{m}_u$. We align the mean lesion confidence across the original and generated regions:

$$\mathcal{L}_{\text{pair}} = \mathbb{E}_{\mathcal{D}_{\text{sym}}} \left[\left(\mu(\mathbf{p}, \mathbf{m}) - \mu(\mathbf{p}, \mathbf{m}^c) \right)^2 \right]. \tag{18}$$

This constrains regional confidence without requiring identical voxel-wise predictions in the two hemispheres.

Overall objective. The complete implantation objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{flip}} \mathcal{L}_{\text{flip}} + \lambda_{\text{pair}} \mathcal{L}_{\text{pair}}. \tag{19}$$

We set $\lambda_{\text{flip}} = 0.1$ and $\lambda_{\text{pair}} = 0.05$. All three regularizers use the highest-resolution output; a regularizer is set to zero when its required samples are absent from the minibatch.