Make Foundation Models Trustworthy Again: Causal Fine-Adaptation for Medical Image Segmentation

Hongpeng Yang[1†], Yingxin Chen[2†], Shiqiang Ma[3*], Fei Guo[2*]

[1]Molinaroli College of Engineering and Computing, University of South Carolina
[2]School of Computer Science and Engineering, Central South University
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
hongpeng@email.sc.edu, chenyingxin@csu.edu.cn, sq.ma@siat.ac.cn, guofei@csu.edu.cn

Abstract

Vision foundation models (e.g., SAM2, CLIP) show strong generalization in natural image analysis but degrade significantly in specialized domains like medical imaging. This is critical for tasks such as brain tumor segmentation, where errors directly affect surgical planning and patient outcomes. In such contexts, segmentation must be highly reliable and structurally precise, underscoring the need for adaptable methods with low error tolerance. While fine-tuning is the dominant strategy, it is computationally expensive and prone to forgetting. To address this, we propose CausalBridgeNet, a causality-guided correction framework for medical image segmentation. Inspired by predictive coding theories of the Bayesian brain, our method introduces a Predictive Causal Reasoning Unit (PCRU) that estimates structured error maps and delivers targeted feedback to iteratively refine predictions. This forms a closed-loop, error-aware correction mechanism without modifying the foundation model. By keeping the backbone frozen, CausalBridgeNet preserves general visual priors while enhancing task-specific accuracy. On the BraTS 2025 benchmark, it achieves an average Dice score of 84.48 and HD95 of 5.48 across tumor subregions, demonstrating its effectiveness for high-precision medical segmentation.

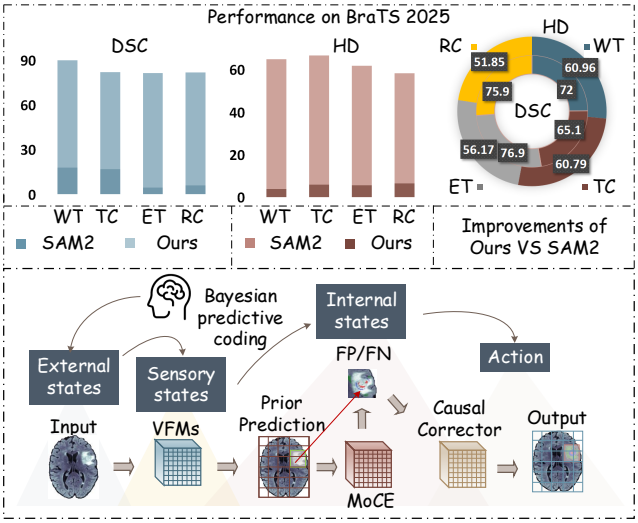

Figure 1: Top: DSC and HD95 improvements over SAM2 on BraTS 2025 across tumor subregions. Bottom: CausalBridgeNet follows a predictive coding principle, where the proposed modules estimates error signals and refines predictions through causal reasoning.

Introduction

Vision foundation models (VFMs), such as CLIP (Radford et al. 2021) and SAM2 (Ravi et al. 2024), have achieved remarkable generalization capabilities in natural image understanding through large-scale pretraining. However, directly applying these models to domain-specific applications like 3D medical image segmentation often results in suboptimal performance. The discrepancy arises from fundamental differences between natural and medical images in terms of modality, structure, and annotation availability. As a result, VFMs tend to exhibit poor inductive bias and limited reliability in high-stakes medical contexts.

A common approach to bridge this gap is to fine-tune VFMs on downstream medical datasets. Yet this strategy is both computationally expensive and prone to catastrophic forgetting, undermining the general knowledge embedded in the pretrained models. Moreover, full fine-tuning compromises the modularity and transferability of VFMs, making them less suitable for plug-and-play deployment in resource-constrained or privacy-sensitive medical settings.

Inspired by the predictive coding theory of the Bayesian brain (Knill and Pouget 2004), which postulates that perception is driven by minimizing structured prediction errors, we propose to view domain adaptation through the lens of causal inference. In this view, the discrepancy between model prediction and input evidence can be interpreted as a structured error signal to guide refinement. Rather than relying on gradient-based fine-tuning of the backbone, we introduce a lightweight, interpretable correction mechanism that explicitly estimates and compensates for systematic prediction errors.

To this end, we propose CausalBridgeNet, a plug-and-play causal adaptation framework that enables vision foun-

*Shiqiang Ma and Fei Guo are corresponding authors.
†These authors contributed equally.

dation models to generalize reliably to 3D medical segmentation tasks. Our method introduces a Predictive Causal Reasoning Unit (PCRU), composed of a Mixture of Causal Experts (MoCE) and a Causal Corrector. The MoCE estimates class-specific false positive and false negative maps guided by learnable prompts, while the corrector refines predictions in a closed-loop fashion—mimicking predictive error minimization in the brain. The backbone remains entirely frozen, making the approach efficient and scalable.

The top of Figure 1 presents quantitative results on the BraTS 2025 benchmark. It compares the performance of our method against directly applying SAM2 to medical images. CausalBridgeNet achieves consistent improvements across all tumor subregions in both Dice and HD95 metrics, demonstrating the effectiveness of causal adaptation over naive VFM transfer. Figure 1 also shows the conceptual illustration of CausalBridgeNet as a Bayesian predictive coding system, where VFM Blocks based U-shape model generates prior predictions (sensory states), causal experts infer structured prediction errors (internal states), and the Causal Corrector applies targeted modulation (action) to refine the output. This loop simulates Bayesian predictive coding to enhance segmentation robustness.

Extensive experiments on medical image segmentation benchmarks demonstrate that CausalBridgeNet significantly improves prediction robustness and generalization without requiring any backbone modification. Our framework offers a practical and principled pathway to adapt vision foundation models for high-risk, domain-specific tasks such as medical image analysis.

In summary, our key contributions are as follows:

- We develop CausalBridgeNet, an efficient and scalable framework that enables VFMs to generalize effectively to 3D medical segmentation tasks without retraining or modifying the VFM backbone.
- We design the Predictive Causal Reasoning Unit (PCRU), which implements a closed-loop causal correction mechanism by estimating structured, class-specific prediction errors and refining outputs. As a fully decoupled module, PCRU supports plug-and-play integration into diverse medical segmentation pipelines with minimal computational cost.
- We conduct extensive experiments on 3D brain tumor and 2D breast tumor segmentation benchmarks. CausalBridgeNet achieves state-of-the-art performance in multiple metrics, demonstrating the practical utility and clinical viability of our causal adaptation paradigm.

Related Works

Vision Foundation Models

Vision foundation models (VFMs), such as CLIP (Radford et al. 2021) and SAM2 (Ravi et al. 2024), have demonstrated impressive generalization capabilities across various natural image tasks, owing to large-scale pretraining on diverse web-scale datasets. By decoupling feature extraction from task-specific training, VFMs provide a reusable backbone for many downstream applications.

However, recent studies (Zhu et al. 2024; Wang et al. 2022) have revealed that the performance of VFMs degrades significantly when applied to domains such as medical imaging, which differ in data modality, structure, and annotation style. This domain discrepancy leads to poor inductive bias and unreliable performance in safety-critical tasks. Existing solutions often rely on full fine-tuning or prompt tuning (Zhou et al. 2022), both of which either require large computational resources or struggle with unstable performance in 3D medical segmentation.

Causal Learning

Causal learning offers a principled approach to handle domain shift and distributional biases by modeling the underlying generative mechanisms behind data (Schölkopf et al. 2021; Luo, Peng, and Ma 2020). In computer vision, causal reasoning has been applied to image segmentation tasks (Li et al. 2024) and to improve model robustness (Miao et al. 2023).

A particularly relevant perspective comes from the predictive coding theory of the Bayesian brain (Friston 2005), which views perception as the iterative minimization of prediction error. This motivates designing architectures that explicitly represent and correct model mispredictions in a closed-loop manner. Inspired by this, we introduce a causal correction pipeline that estimates structured false positive and false negative maps and integrates them into downstream predictions through learnable class-specific updates.

Mixture of Experts

Mixture of Experts (MoE) models (Jacobs et al. 1991) aim to increase model capacity by routing inputs to specialized expert subnetworks. MoE architectures have achieved state-of-the-art performance in language modeling and vision-language tasks by dynamically activating a subset of experts based on learned gating functions (Zoph et al. 2022; Riquelme et al. 2021).

In medical image segmentation, Luo et al. (Luo et al. 2025) proposed TA-MoSC, a task-adaptive MoE-based module that dynamically selects skip connections in U-Net via a learnable router. Their method improves generalization across diverse datasets with minimal overhead.

Method

Overview

We propose CausalBridgeNet, a general-purpose causal correction framework that improves the cross-domain generalization of frozen VFMs, such as CLIP and SAM2, in high-stakes domains like medical image segmentation. Unlike conventional fine-tuning approaches, CausalBridgeNet leaves the backbone model untouched and introduces a interpretable correction unit to refine predictions in a biologically inspired manner.

As illustrated in Figure 2, it first produces coarse predictions via a frozen VFM-based U-shape pipeline. Then, a Predictive Causal Reasoning Unit (PCRU) estimates structured prediction errors using a Mixture of Causal Experts

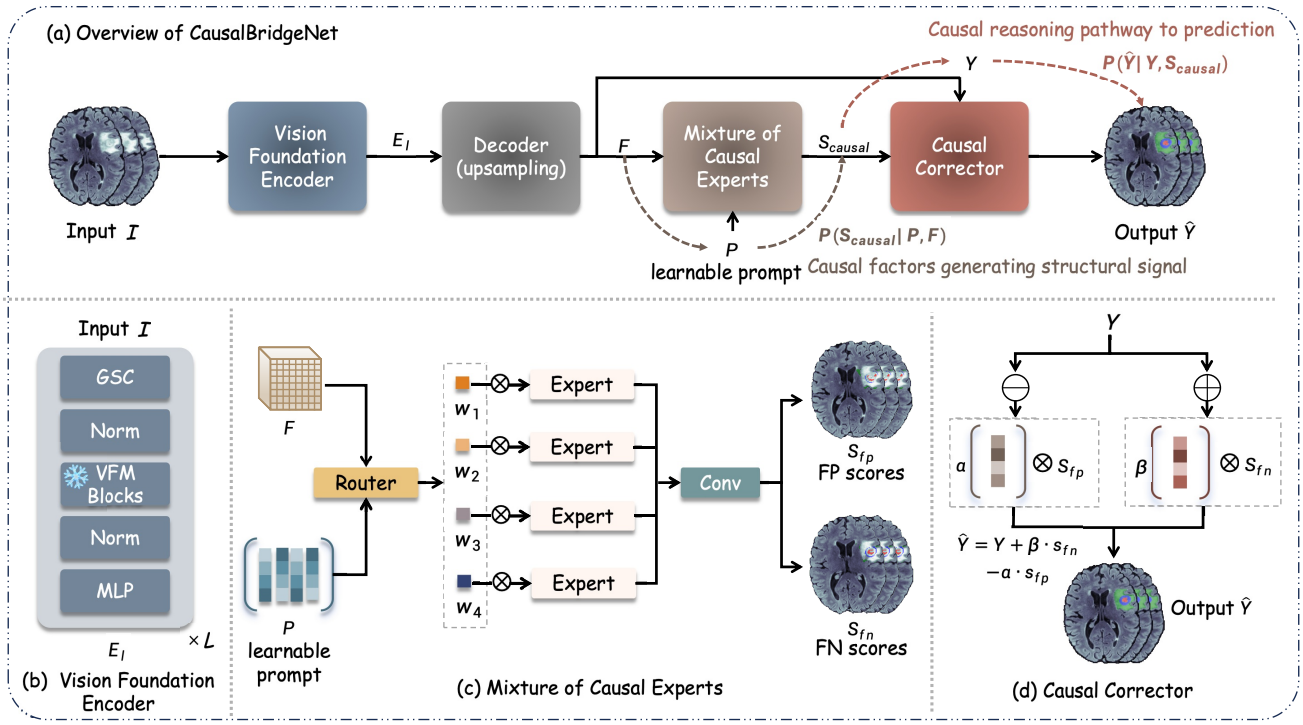

Figure 2: Overview of CausalBridgeNet. (a) Overall pipeline: initial predictions from frozen VFMs are refined via a causal reasoning loop. (b) Encoder integrates frozen VFM blocks with adapters. (c) Mixture of Causal Experts estimates structured error maps using learnable prompt. (d) Causal Corrector applies class-wise adjustments to improve segmentation accuracy.

(MoCE) module guided by contextual prompts. These error estimates are subsequently used to refine the predictions through a class-aware correction mechanism Causal Corrector. This design enables efficient, interpretable, and plug-and-play adaptation of large vision models to new domains, without incurring catastrophic forgetting or heavy computational overhead.

Vision Foundation Encoder

To adapt vision foundation models (VFMs), such as CLIP and SAM, to 3D volumetric medical images, we construct a hierarchical encoder that integrates frozen VFM blocks into a U-shaped architecture. This encoder progressively downsamples the input volume while preserving semantic richness across multiple scales. To support volumetric processing, we incorporate lightweight modules including Gated Spatial Convolution (GSC) (Xing et al. 2024), layer normalization, and MLP-based channel adapters.

Given an input image $I \in \mathbb{R}^{C\times D\times H\times W}$, we first apply a 3D stem layer to produce the initial low-level feature map E_0. This output is then processed through a sequence of adaptive blocks across four hierarchical stages. At each stage $l \in \{1, 2, 3, 4\}$, the feature is computed as:

$$\hat{E}_l = \text{GSC}(E_{l-1}) \tag{1}$$

$$\tilde{E}_l = \text{VFB}(\text{LN}(\hat{E}_l)) \tag{2}$$

$$E_l = \text{MLP}(\text{LN}(\tilde{E}_l)) \tag{3}$$

Where $\text{GSC}(\cdot)$ denotes the Gated Spatial Convolution module, $\text{VFB}(\cdot)$ represents grouped transformer blocks from the frozen vision foundation model, $\text{LN}(\cdot)$ is layer normalization, and $\text{MLP}(\cdot)$ is a channel-wise projection module to enrich feature representations.

To support hierarchical learning, each stage is followed by a downsampling operation, such that the resolution of the input volume $D \times H \times W$ is progressively reduced to $D/16 \times H/16 \times W/16$, similar to UNETR (Hatamizadeh et al. 2022).

The resulting multi-scale features $\{E_1, E_2, E_3, E_4\}$ capture increasingly abstract semantic information and are passed to the decoder and causal correction modules for downstream prediction and refinement.

Semantic Decoder

The Semantic Decoder reconstructs dense segmentation predictions by progressively upsampling the multi-scale encoder features $\{E_1, E_2, E_3, E_4\}$ using a U-shaped architecture. At each decoder stage, high-level features are upsampled and fused with the corresponding encoder output via skip connections, allowing fine-grained spatial information to be recovered.

Let D_4 denote the bottleneck feature initialized as $D_4 = E_4$. Then, for each stage $l \in \{3, 2, 1\}$, we compute the decoder output by first upsampling the deeper representation and then combining it with the encoder feature E_l:

Models	WT		TC		ET		RC	
	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
UNETR(Hatamizadeh et al. 2022)	85.95	8.21	71.75	11.17	72.33	10.01	51.05	22.44
SwinUNETR (Hatamizadeh et al. 2021)	89.42	5.63	75.81	9.30	74.99	8.93	67.27	14.62
SegResNet (Myronenko 2018)	89.43	5.36	75.70	8.68	74.86	8.42	74.91	10.09
SegMamba (Xing et al. 2024)	90.02	4.75	78.20	7.79	77.20	7.50	76.16	10.25
CausalBridgeNet(CLIP)	<u>91.25</u>	<u>4.47</u>	<u>82.11</u>	**5.80**	<u>81.51</u>	**5.44**	<u>81.78</u>	<u>6.94</u>
CausalBridgeNet(SAM2)	**91.37**	**3.90**	**82.72**	<u>5.90</u>	**81.72**	<u>5.61</u>	**82.12**	**6.49**

Table 1: Comparison of segmentation performance across different models on BraTS 2025. Metrics include DSC (higher is better) and HD95 (lower is better) for WT, TC, ET, and RC.

$$\tilde{D}_l = \text{UpConv}_l(D_{l+1}) \tag{4}$$

$$D_l = \text{DecoderBlock}_l([\tilde{D}_l, E_l]) \tag{5}$$

Here, $\text{UpConv}_l(\cdot)$ denotes a 3D transposed convolution or upsampling layer at level l, and $[\cdot,\cdot]$ denotes channel-wise concatenation. $\text{DecoderBlock}_l(\cdot)$ is a convolutional residual block that fuses and refines the combined features.

After the final decoder stage D_1, we apply an additional convolutional head to produce the raw segmentation logits:

$$Y = \text{Conv}_{\text{out}}(D_1) \tag{6}$$

where $Y \in \mathbb{R}^{C_{\text{seg}} \times D \times H \times W}$ and C_{seg} is the number of target classes.

This decoding strategy enables efficient multi-scale feature fusion while preserving boundary precision and fine structural details, which is critical for high-resolution 3D medical segmentation tasks.

Predictive Causal Reasoning Unit (PCRU)

Inspired by predictive coding in the Bayesian brain, PCRU introduces an explicit causal correction loop to refine predictions based on systematic error patterns. It comprises two components: (1) a Mixture of Causal Experts that estimates false positive/negative maps, and (2) a Causal Corrector that adjusts the logits accordingly.

Mixture of Causal Experts To explicitly model and correct systematic prediction errors caused by domain shift, we introduce a Mixture of Causal Experts (MoCE). This module estimates structured false positive (FP) and false negative (FN) maps conditioned on the intermediate decoder features and semantic predictions, enabling fine-grained causal refinement.

Given the intermediate decoder feature map $F \in \mathbb{R}^{C_f \times D \times H \times W}$ and the coarse segmentation logits $Y \in \mathbb{R}^{C_{\text{seg}} \times D \times H \times W}$, we first concatenate F with a learned 3D prompt $P \in \mathbb{R}^{C_p \times D \times H \times W}$ to form the adapted input:

$$\tilde{F} = \text{Concat}(F, P) \tag{7}$$

The fused feature $\tilde{F}$ is passed to a Router module, which computes expert selection weights based on global context. Specifically, the Router applies global average pooling (GAP) over spatial and depth dimensions, followed by a linear projection and softmax normalization:

$$\mathbf{w} = \text{Softmax}(\text{FC}(\text{GAP}(\tilde{F}))) \tag{8}$$

Where $\mathbf{w} \in \mathbb{R}^K$ denotes the mixture weights assigned to the K causal experts. The GAP operation reduces $\tilde{F} \in \mathbb{R}^{C \times D \times H \times W}$ to a global descriptor in $\mathbb{R}^C$.

Each expert $\mathcal{E}_k$ independently transforms $\tilde{F}$, and the final expert output is obtained as a weighted sum:

$$S_{\text{causal}} = \sum_{k=1}^{K} w_k \cdot \mathcal{E}_k(\tilde{F}) \tag{9}$$

where $S_{\text{causal}} \in \mathbb{R}^{2C_{\text{seg}} \times D \times H \times W}$ contains the estimated error signals.

We interpret S_{causal} as a concatenation of false positive and false negative score maps for each class:

$$S_{\text{causal}} = [S_{\text{fp}} \,\|\, S_{\text{fn}}], \quad S_{\text{fp}}, S_{\text{fn}} \in \mathbb{R}^{C_{\text{seg}} \times D \times H \times W} \tag{10}$$

These error maps are passed to the causal correction module to refine the prediction logits. Since only the adapter components are trainable, the method is both memory-efficient and compatible with any frozen encoder-decoder backbone.

Causal Corrector To integrate the estimated false positive (FP) and false negative (FN) signals into the final prediction, we introduce a class-wise Causal Corrector. This module applies interpretable, additive adjustments to the coarse segmentation logits based on the error maps produced by the Causal Adapter.

Let $Y \in \mathbb{R}^{C_{\text{seg}} \times D \times H \times W}$ denote the coarse logits from the decoder, and let $S_{\text{fp}}, S_{\text{fn}} \in \mathbb{R}^{C_{\text{seg}} \times D \times H \times W}$ denote the predicted false positive and false negative maps for each class, respectively. The refined logits $\hat{Y}$ are computed as:

$$\hat{Y}_c = Y_c + \beta_c \cdot S^c_{\text{fn}} - \alpha_c \cdot S^c_{\text{fp}}, \quad \forall c \in \{1, \ldots, C_{\text{seg}}\} \tag{11}$$

where α_c and β_c are learnable scalar weights that modulate the correction strength for each class c.

This correction formulation enables class-aware, direction-sensitive refinement: increasing logit confidence in regions prone to false negatives and suppressing false positives accordingly. Since α_c and β_c are shared across the volume, the module introduces minimal overhead while maintaining interpretability.

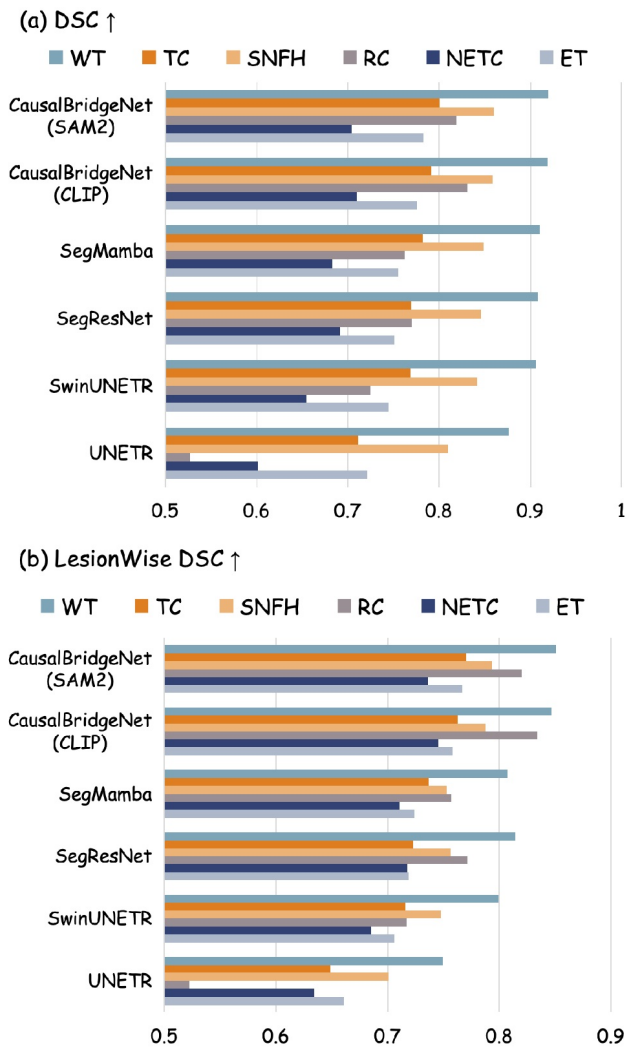

Figure 3: (a) Standard and (b) lesion-wise DSC scores on the BraTS2025 online validation set

Experiments

Datasets and Evaluation Metrics

We evaluate the proposed CausalBridgeNet on three datasets: BraTS2025 Pre-only, BraTS2025 Pre+Post, and BUSI. This diverse setup enables rigorous assessment of the model's generalization under domain shifts and structural variability.

BraTS2025 Pre-only This dataset is derived from the 2025 BraTS Glioma Segmentation Challenge and includes only the pre-treatment MRI scans of patients with diffuse gliomas. Each volume contains four modalities: T1, T1ce, T2, and FLAIR. Annotations are provided for three tumor subregions: *enhancing tumor (ET)*, *tumor core (TC)*, and *whole tumor (WT)*. The dataset is divided into `TrainingData` (1,251 cases) and `ValidationData` (219 cases).

BraTS2025 Pre+Post This configuration includes both pre-treatment and post-treatment volumes with full annotation coverage. The dataset is partitioned into `TrainingData` containing 1,251 pre-treatment and 1,567 post-treatment cases (total 2,818), and `ValidationData` containing 219 pre-treatment and 188 post-treatment cases (total 407). All four tumor-related structures—ET, TC, WT, and *resection cavity (RC)*—are included in this settings.

BUSI The Breast Ultrasound Images (BUSI) dataset (Al-Dhabyani et al. 2020) contains 647 grayscale ultrasound images annotated for breast tumor segmentation. We use images labeled as *benign* or *malignant*, excluding normal samples. All images are resized to 256×256 pixels. This dataset presents distinct challenges such as low contrast, speckle noise, and variable tumor morphology, making it well-suited for testing cross-modal generalization.

Evaluation Metrics We adopt evaluation protocols tailored to each dataset and experiment configuration to ensure fair and task-appropriate comparison.

- **BraTS2025 Pre-only and Pre+Post:** Following the experimental protocol of SegMamba (Xing et al. 2024), we split those datasets into training/validation/testing sets using a 70%/10%/20% ratio. Evaluation is conducted on the held-out internal test set using two volumetric metrics: the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95), computed for each tumor subregion.
- **BraTS2025 ValidationData:** We submit predictions on the `ValidationData` to the official online BraTS2025 evaluation platform. It provides both standard and Lesion-wise DSC as segmentation metrics.
- **BUSI:** For the BUSI ultrasound dataset, we follow the evaluation setting in U-KAN (Li et al. 2025). The dataset was randomly split into 80% training and 20% validation subsets. Results are reported over three random runs. As the task involves 2D segmentation of benign and malignant tumors, we adopt Intersection over Union (IoU) and F1 as the metrics.

Implementation Details

For the BraTS2025 dataset (both Pre-only and Pre+Post settings), we follow the 3D training pipeline established in SegMamba (Xing et al. 2024). Our model uses CLIP and SAM2 as frozen vision foundation backbones throughout all experiments. We adopt a 3D crop size of $(64 \times 64 \times 64)$ and a batch size of 2. The model is optimized using a combination of Dice loss and cross-entropy loss. We use a stochastic gradient descent (SGD) optimizer and a polynomial learning rate scheduler (initial learning rate of 0.01 and decay of 1×10^{-5}). Training runs for 1,000 epochs with data augmentations including brightness, gamma, rotation, scaling, mirror, and elastic deformation.

For the BUSI dataset, we follow the training strategy from U-KAN (Li et al. 2025). We use a batch size of 8 and an initial learning rate of 1×10^{-4}, optimized with Adam and scheduled using cosine annealing (minimum learning rate 1×10^{-5}). The loss function is a weighted sum of binary

Models	WT		TC		ET		Avg	
	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
UNETR(Hatamizadeh et al. 2022)	91.93	5.64	89.36	5.43	83.69	5.85	88.33	5.64
SwinUNETR (Hatamizadeh et al. 2021)	93.11	5.02	91.73	4.02	85.52	4.48	90.12	4.51
SegResNet (Myronenko 2018)	93.18	4.29	91.30	3.91	86.39	4.16	90.29	4.12
SegMamba (Xing et al. 2024)	93.32	4.67	**91.90**	4.27	86.69	4.53	90.64	4.49
CausalBridgeNet (CLIP)	<u>93.75</u>	<u>3.43</u>	91.39	**3.49**	**87.88**	<u>3.84</u>	<u>91.01</u>	<u>3.59</u>
CausalBridgeNet (SAM2)	**93.83**	**3.40**	<u>91.80</u>	<u>3.52</u>	<u>87.85</u>	**3.82**	**91.16**	**3.58**

Table 2: Comparison of segmentation performance across different models on BraTS 2025 Pre-only dataset. Metrics include DSC and HD95 for WT, TC, ET, and Average.

cross-entropy (BCE) and Dice loss. Vanilla data augmentations (random rotation and flipping) are applied. The model is trained for 400 epochs, and results are averaged over three random seeds to ensure stability.

Evaluation Results

Results on Pre+Post dataset Table 1 reports the segmentation performance of different models on the BraTS 2025 dataset. Our proposed CausalBridgeNet achieves the best performance across all tumor structures, consistently outperforming previous models, including UNETER (Hatamizadeh et al. 2022), SwinUNETR (Hatamizadeh et al. 2021), SegResNet (Myronenko 2018), and SegMamba (Xing et al. 2024). Specifically, CausalBridgeNet (SAM2) achieves a DSC of 82.72 on TC and 82.12 on RC, surpassing SegMamba by +4.52 and +5.96 points, respectively. The HD95 scores also show significant improvement, indicating better boundary accuracy.

Notably, both CausalBridgeNet variants (with CLIP and SAM2 backbones) outperform conventional encoder-decoder architectures, demonstrating the effectiveness of our causal correction mechanism and the benefits of leveraging frozen vision foundation models. Even with the backbone is not specifically pretrained on medical images, our model yields substantial gains, suggesting strong generalization capabilities.

Online Validation Results. We evaluate CausalBridgeNet on the BraTS2025 official online validation platform using the Pre+Post training configuration. Figure 3 illustrate the segmentation performance using both standard and lesion-wise metrics across six tumor-related regions: ET, NETC, RC, SNFH, TC, and WT.

As shown in Figure 3, CausalBridgeNet achieves consistently superior DSC performance across both global and lesion-level evaluations. CausalBridgeNet (SAM2) achieves the highest DSC scores on ET (78.26), TC (80.05), and WT (91.96), and also leads in lesion-wise DSC on WT (85.05), demonstrating strong generalization on whole tumor structures and enhancing clinical reliability. CausalBridgeNet (CLIP) outperforms all models on RC (83.11) and NETC (70.99) in DSC, and achieves the highest lesion-wise DSC on RC (83.39), ET (75.79), and NETC (74.54), reflecting its strength in segmenting core tumor components. Both variants clearly outperform competitive baselines across nearly all tumor subregions. These results highlight the effectiveness of our causal reasoning mechanism in leveraging frozen vision backbones to deliver robust and precise 3D medical segmentation.

Results on Pre-only Dataset Table 2 presents the segmentation results on the BraTS2025 Pre-only dataset, where only pre-treatment scans and three tumor subregions (WT, TC, ET) are considered. As shown, CausalBridgeNet (SAM2) achieves the highest average DSC score of 91.16 and the lowest average HD95 of 3.58, outperforming all baseline models. Specifically, our method demonstrates superior performance in segmenting the enhancing tumor (ET), achieving a HD95 of 3.82, compared to 4.53 by SegMamba. This indicates improved boundary precision in regions known for low contrast and complex morphology.

CausalBridgeNet with the CLIP backbone also achieves competitive performance, with an average DSC of 91.01, surpassing all fully-trained baselines despite using a frozen vision backbone not pretrained on medical data. These results reinforce the effectiveness of our causal refinement strategy in boosting segmentation accuracy.

Ablation Studies

To assess the effectiveness and generalizability of the proposed PCRU, we conduct two sets of ablation studies: (1) removing the module from CausalBridgeNet on the BraTS2025 dataset, and (2) integrating it into U-KAN and evaluating on the BUSI dataset.

Impact on CausalBridgeNet (BraTS2025). Table 3 reports the segmentation results of CausalBridgeNet (SAM2) with and without the PCRU. Removing the module leads to consistent performance drops across all four tumor subregions. For example, Dice scores decrease by 1.79 points on TC (82.72 $\rightarrow$ 80.93) and 1.56 points on RC (82.12 $\rightarrow$ 80.26). Likewise, HD95 increases notably in RC (6.49 $\rightarrow$ 7.32), indicating degraded boundary accuracy. These results highlight the importance of our module in refining coarse predictions and correcting semantic misalignment, particularly in anatomically irregular regions.

Transfer to U-KAN (BUSI). To further evaluate the generalizability of the causal module, we incorporate PCRU into U-KAN (Li et al. 2025) and test on the BUSI ultrasound dataset. As shown in Table 4, the extended U-KAN with

Models	WT		TC		ET		RC	
	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
CausalBridgeNet (SAM2) w/o PCRU	90.74	4.98	80.93	6.60	80.17	6.12	80.26	7.32
CausalBridgeNet (SAM2)	**91.37**	**3.90**	**82.72**	**5.90**	**81.72**	**5.61**	**82.12**	**6.49**

Table 3: Ablation study comparing the segmentation performance of CausalBridgeNet with and without the PCRU, using SAM2 as backbone. Metrics include DSC score and HD95 for four tumor subregions.

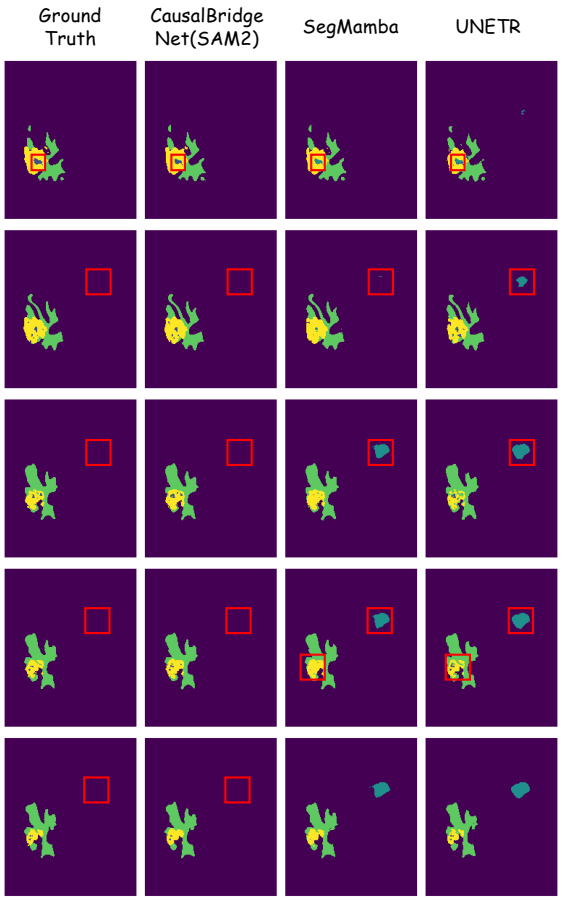

Figure 4: Visualized segmentation results. Each column corresponds to a different model's prediction and label; each row shows a different case from the BraTS2025 dataset.

our correction module significantly outperforms all baselines, improving IoU from 63.38% to 68.85%, and F1-score from 76.40% to 80.86%. This demonstrates that the proposed module is architecture-agnostic and remains effective in low-contrast, cross-modal settings.

Visualization

To further verify the effectiveness of our approach, we present qualitative comparisons in Figure 4. It presents segmentation outputs on four representative cases from the BraTS 2025 dataset. The first row shows a small tumor region where both SegMamba and UNETR misclassify the tumor class, leading to semantic errors in prediction; in contrast, CausalBridgeNet (SAM2) preserves the class identity and shape more faithfully. In the second and third rows, baseline models fail to recover fine-grained boundaries or exhibit partial tumor omission, particularly in areas with irregular morphology.

Notably, CausalBridgeNet produces spatially coherent predictions with reduced false negatives. In the fourth and fifth rows, although the tumor region is large, SegMamba and UNETR introduce scattered false positives (blue region), while our method maintains clearer boundaries and stronger alignment with ground truth. These visual results confirm that the causal refinement pathway effectively corrects structural inconsistencies and improves segmentation fidelity in challenging clinical cases.

Models	IoU ↑	F1 ↑
U-Net (Ronneberger, Fischer, and Brox 2015)	57.22±4.74	71.91±3.54
Att-Unet (Oktay et al. 2018)	55.18±3.61	70.22±2.88
U-Net++ (Zhou et al. 2018)	57.41±4.77	72.11±3.90
U-NeXt (Valanarasu and Patel 2022)	59.06±1.03	73.08±1.32
Rolling-UNet (Liu et al. 2024)	61.00±0.64	74.67±1.24
U-Mamba (Ma, Li, and Wang 2024)	61.81±3.24	75.55±3.01
U-KAN (Li et al. 2025)	63.38±2.83	76.40±2.90
Ours	**68.85±2.01**	**80.86±1.64**

Table 4: Evaluation on BUSI: comparison of U-KAN with our PCRU and baselines. Metrics are averaged over three runs.

Conclusion

In this work, we propose CausalBridgeNet, a novel framework that integrates vision foundation models with a causal correction mechanism to address domain shifts and error accumulation in medical image segmentation. By bridging frozen VFMs backbone with U-shaped network through a learnable refinement module, our method enables both robust representation transfer and error-aware correction. Extensive experiments on BraTS2025 demonstrate that CausalBridgeNet achieves state-of-the-art performance across both standard and lesion-wise evaluations. Ablation studies confirm the efficacy of the causal correction module, while additional transfer experiments on the BUSI dataset show strong generalizability across modalities and architectures. Our findings suggest that causality-guided correction, when coupled with powerful frozen vision models, offers a promising direction for building modular, reliable, and cross-domain segmentation systems in medical imaging.

Acknowledgments

This work is supported by grants from the National Natural Science Foundation of China (NSFC 62322215, 62532017, 62402488). This study was also supported in part by the High-Performance Computing Center of Central South University.

References

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in brief*, 28: 104863.

Friston, K. 2005. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456): 815–836.

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H. R.; and Xu, D. 2021. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, 272–284. Springer.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 574–584.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Knill, D. C.; and Pouget, A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12): 712–719.

Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Liu, Y.; Chen, Z.; and Yuan, Y. 2025. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4652–4660.

Li, C.; Mao, Y.; Liang, S.; Li, J.; Wang, Y.; and Guo, Y. 2024. Deep causal learning for pancreatic cancer segmentation in CT sequences. *Neural Networks*, 175: 106294.

Liu, Y.; Zhu, H.; Liu, M.; Yu, H.; Chen, Z.; and Gao, J. 2024. Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 3819–3827.

Luo, Y.; Peng, J.; and Ma, J. 2020. When causal inference meets deep learning. *Nature Machine Intelligence*, 2(8): 426–427.

Luo, Z.; Zhu, X.; Zhang, L.; and Sun, B. 2025. Rethinking U-Net: Task-Adaptive Mixture of Skip Connections for Enhanced Medical Image Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5874–5882.

Ma, J.; Li, F.; and Wang, B. 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*.

Miao, J.; Chen, C.; Liu, F.; Wei, H.; and Heng, P.-A. 2023. Caussl: Causality-inspired semi-supervised learning for medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 21426–21437.

Myronenko, A. 2018. 3D MRI brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, 311–320. Springer.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634.

Valanarasu, J. M. J.; and Patel, V. M. 2022. Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention*, 23–33. Springer.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, 3876.

Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; and Zhu, L. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 578–588. Springer.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, 3–11. Springer.

Zhu, J.; Hamdi, A.; Qi, Y.; Jin, Y.; and Wu, J. 2024. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*.

Zoph, B.; Bello, I.; Kumar, S.; Du, N.; Huang, Y.; Dean, J.; Shazeer, N.; and Fedus, W. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.