

Article

A Mathematical Theory of Phase-Consistent Information Bottleneck for Cross-Domain Generalization

Feng Liu *  and Zheng Wang *

School of Information Engineering, Shandong Youth University of Political Science, Jinan 250103, China

* Correspondence: liusdyu@sina.com (F.L.); wangzheng@sdyu.edu.cn (Z.W.)

Abstract

We propose a mathematical framework for domain generalization in medical image segmentation built on dual-tree complex wavelet transform (DTCWT) and variational information theory. The core premise is that, under adequate spatial normalization and acquisition-style shifts, DTCWT phase components are more closely associated with anatomical structure, whereas amplitude components are more sensitive to domain-specific intensity and style variations. We formulate this as a local phase–magnitude complementarity premise and construct an information bottleneck that operates on structured subband representations. The framework provides several key theoretical results under explicit structural assumptions: an information bound showing when DTCWT amplitude subbands better isolate domain-related information than global Fourier representations; a variational information bottleneck encoder that compresses domain-specific amplitude information into low-dimensional latent codes; a triple constraint mechanism (domain supervision, KL compression, and orthogonality) that controls domain–task information leakage; and a predictive feature modulation scheme with $O(1)$ spatial complexity. We further analyze test-time adaptation via calibrated uncertainty, deriving a sufficient condition under which a two-pass inference strategy reduces the expected generalization gap. Finally, we include illustrative public-dataset checks on FeTS 2022 and BraTS 2023 to test the central phase–amplitude premise and the feasibility of DTCWT-front-end segmentation. All theorems are stated with their assumptions and verifiable conditions, offering a physically motivated approach to domain generalization in medical imaging.

Keywords: information bottleneck; dual-tree complex wavelet transform; domain generalization; disentangled representation; phase consistency; variational inference; medical image segmentation



Academic Editor: Andrea Murari

Received: 22 May 2026

Revised: 27 June 2026

Accepted: 29 June 2026

Published: 3 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Medical image segmentation models are commonly trained on data collected from a limited number of hospitals, scanners, and acquisition protocols. When such models are deployed at a new institution, the anatomical structures of interest may remain clinically comparable, but image contrast, bias fields, noise statistics, reconstruction artifacts, and intensity normalization can change substantially. This mismatch is a practical obstacle for brain tumor, organ, and lesion segmentation: a model can learn accurate boundaries on the source cohort yet fail on a target site because the same tissue class appears with different local intensity and texture characteristics.

The central question of this paper is therefore not only how to improve a network architecture, but how to describe domain shift in a representation where the problem

becomes analyzable. We use the dual-tree complex wavelet transform (DTCWT) as such a representation. The working premise is that, after spatial normalization, local phase is more closely tied to anatomical structure, whereas local amplitude is more sensitive to acquisition style and scanner-dependent appearance. The rest of the paper develops a conditional information-theoretic framework around this premise, making explicit where standard tools are used and where new theoretical claims are introduced.

1.1. The Domain Generalization Problem

Let $(\mathcal{X}, \mathcal{Y})$ denote the input and label spaces. We consider K source domains $\mathcal{D}_S = \{\mathcal{D}_1, \dots, \mathcal{D}_K\}$, where each domain

$$\mathcal{D}_k = \{(\mathbf{x}_i^k, \mathbf{y}_i^k)\}_{i=1}^{N_k}, \quad \mathbf{x}_i^k \in \mathcal{X}, \mathbf{y}_i^k \in \mathcal{Y} \quad (1)$$

is drawn from an unknown joint distribution $P_k(\mathbf{x}, \mathbf{y})$. The domain generalization objective is to learn a hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes the risk on an unseen target domain $\mathcal{D}_T \sim P_T$:

$$R_T(h) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim P_T}[\ell(h(\mathbf{x}), \mathbf{y})] \quad (2)$$

using only samples from $\mathcal{D}_S$, without any access to target labels.

The following prior-work bound is used only as theoretical motivation. The classical Ben-David et al. generalization bound [1] provides the starting point:

$$R_T(h) \leq R_S(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(P_S, P_T) + \lambda^* \quad (3)$$

where $d_{\mathcal{H}\Delta\mathcal{H}}$ is the $\mathcal{H}\Delta\mathcal{H}$ -divergence between source and target distributions, and λ^* is the combined optimal error.

1.2. The Frequency-Domain Structure of Domain Shift

A key structural observation, dating back to Oppenheim and Lim [2], is that in many signal processing contexts, the Fourier phase spectrum carries structural information while the magnitude spectrum carries intensity and style information. We extend this intuition to the medical imaging domain as a structural assumption whose validity depends on spatial normalization and on domain shifts that primarily affect acquisition appearance rather than gross anatomy.

Consider an invertible frequency-domain transform $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Z}$, where $\mathcal{Z}$ is a complex-valued space admitting polar decomposition $\mathcal{F}(\mathbf{x}) = \mathbf{A} \odot e^{i\Phi}$ with amplitude $\mathbf{A}$ and phase Φ .

Property 1 (Conditional Frequency-Domain Domain Shift). *For two domains $\mathcal{D}_k, \mathcal{D}_l$ whose shift is dominated by acquisition appearance rather than gross anatomical change, the following approximate relationships are expected to hold under spatial normalization:*

$$|\mathcal{F}(\mathbf{x}^k)|(u, v) \neq |\mathcal{F}(\mathbf{x}^l)|(u, v) \quad (\text{amplitude varies across domains}) \quad (4)$$

$$\angle \mathcal{F}(\mathbf{x}^k)(u, v) \approx \angle \mathcal{F}(\mathbf{x}^l)(u, v) \quad (\text{phase is approximately invariant}) \quad (5)$$

Equation (4) captures the observation that MRI contrast, gain, and intensity distributions can differ systematically across institutions. Equation (5) reflects the weaker and explicitly conditional premise that anatomical structures—organ boundaries, lesion contours, tissue interfaces—are more stable than acquisition appearance after spatial normalization.

1.3. Limitations of Global Fourier Approaches

Existing frequency-domain domain generalization methods [3,4] operate on the global Fourier transform $\mathcal{F}_{\text{FFT}}$. While computationally convenient, global FFT suffers from a fundamental tension:

Proposition 1 (Spatial Locality Failure of Global FFT). *Let $\mathbf{x} \in \mathbb{R}^{H \times W}$ and let $\tilde{\mathbf{x}}$ be a spatially localized modification of $\mathbf{x}$ on a compact region $\Omega \subset \text{supp}(\mathbf{x})$ with $|\Omega| \ll HW$. Under the global FFT amplitude modification $\mathbf{A}_{\text{FFT}} \mapsto \tilde{\mathbf{A}}_{\text{FFT}}$ performed on the entire frequency plane, the perturbation satisfies*

$$\|\tilde{\mathbf{x}} - \mathbf{x}\|_F^2 = \|\tilde{\mathbf{A}}_{\text{FFT}} \odot e^{i\Phi_{\text{FFT}}} - \mathbf{A}_{\text{FFT}} \odot e^{i\Phi_{\text{FFT}}}\|_F^2 = \|\tilde{\mathbf{A}}_{\text{FFT}} - \mathbf{A}_{\text{FFT}}\|_F^2 \quad (6)$$

where the Parseval equality implies that the spatial-domain perturbation energy is identical to the full-spectrum amplitude perturbation, regardless of spatial locality.

This means: modifying any frequency bin affects every spatial pixel. Domain shifts in MRI are often spatially heterogeneous (e.g., frontal lobe brighter but occipital lobe normal in a particular scanner), and global FFT cannot isolate these local effects.

1.4. The Conditional Local Phase-Magnitude Complementarity Premise

The DTCWT [5] provides a solution through its multi-scale, multi-directional decomposition with approximate shift invariance. Its construction builds on complex and multidimensional dual-tree wavelets and the broader foundations of wavelet analysis [6–9]. We formalize the core physical insight:

Property 2 (Conditional Local Phase-Magnitude Complementarity in Medical Imaging). *Let $\mathbf{x}^k, \mathbf{x}^l$ be MRI images of the same anatomical structure acquired at different institutions. When the dominant cross-domain differences are intensity, contrast, bias-field, or reconstruction effects rather than anatomical deformation, the DTCWT $\mathcal{W}$ at scale j and orientation θ is assumed to satisfy:*

$$\angle \mathcal{W}(\mathbf{x}^k)_{j,\theta} \approx \angle \mathcal{W}(\mathbf{x}^l)_{j,\theta} \quad (7)$$

$$\|\mathcal{W}(\mathbf{x}^k)_{j,\theta}\| = \mathcal{T}_{k \rightarrow l}(\|\mathcal{W}(\mathbf{x}^l)_{j,\theta}\|) \quad (8)$$

where $\mathcal{T}_{k \rightarrow l}$ is a spatially local, generally nonlinear energy redistribution operator acting on the amplitude manifold.

Property 2 is the mathematical cornerstone of our framework. Under this premise, domain shift is structured in the DTCWT domain: phase is treated as a relatively anatomy-associated component, whereas amplitude is treated as an acquisition-sensitive and spatially localized component. It should be read as a conditional structural premise, not as an unconditional claim about all medical images. It is expected to hold best after spatial registration or resampling, and may fail under severe misregistration, major anatomical deformation, missing modalities, or cross-modality shifts where the imaging physics changes the structural signal itself.

1.5. Scope, Evidence, and Falsifiability of the Premise

Because subsequent results depend on Property 2, we explicitly separate three levels of claim. First, the premise is physically motivated: MRI site shifts often arise from acquisition-dependent contrast, gain, bias-field, and reconstruction differences, whereas anatomical boundaries are expected to be more stable after spatial normalization. Second, the premise is representation-specific: it is not a statement about raw pixels or global Fourier coefficients,

but about local complex-wavelet coefficients whose phase and amplitude are spatially indexed. Third, the premise is empirically falsifiable: if DTCWT phase statistics predict acquisition site as strongly as, or more strongly than, amplitude statistics, or if DTCWT front ends fail to support non-trivial segmentation on public data, then the proposed decomposition would not be a useful basis for the theory.

The manuscript therefore treats the theorems below as conditional results. The empirical checks in Section 11 are not intended to prove a universal physical law; they test whether the premise is plausible in representative multi-site brain-tumor MRI settings. This distinction is important for interpretation: the theory provides sufficient conditions and failure-mode analysis, while the experiments assess whether the assumed conditions are reasonable enough to motivate further methodological development.

1.6. Contributions

This work makes several theoretical contributions. First, we formalize a conditional phase–magnitude domain structure under DTCWT and establish a bound on information loss introduced by axial-plane approximation. Second, we propose a variational information bottleneck tailored to structured DTCWT subbands and analyze how subband representations reduce the effective dimensionality of information compression compared to raw spatial inputs. Third, we derive a triple-constraint analysis for disentanglement, showing that domain supervision, KL compression, and orthogonality address distinct failure modes of task/domain separation. Fourth, we design a predictive modulation strategy that preserves information bottleneck semantics while lowering spatial parameter complexity to $O(1)$ with respect to the amplitude volume size. Fifth, we present a theoretical analysis of uncertainty-guided test-time adaptation, deriving a sufficient condition for generalization gap reduction by a two-pass inference procedure. Finally, we add illustrative public-dataset validation to test the phase–amplitude premise and demonstrate DTCWT-front-end segmentation feasibility without reframing the paper as a full benchmark study.

2. Related Work

2.1. Frequency-Domain Domain Generalization

Frequency-domain methods have been used to reduce domain shift by manipulating Fourier amplitude while preserving Fourier phase. Fourier Domain Adaptation (FDA) replaces low-frequency amplitude components between source and target images for semantic segmentation [3], and later Fourier-based domain generalization work extends this idea without requiring target labels [4,10]. These methods motivate the use of phase–amplitude structure, but their global Fourier basis does not localize domain shifts in space. Our framework differs by using DTCWT subbands, which provide multiscale and multidirectional localization.

2.2. Information Bottleneck and Information-Theoretic DG

The information bottleneck principle [11] and its variational implementation [12] provide the compression language used in this paper. Related studies have examined information-theoretic representations, noisy compression, and the limits of information-bottleneck explanations in deep learning [13–15]. Recent information-theoretic domain generalization methods are directly relevant. Invariant information bottleneck (IIB) formulates domain generalization through invariant causal prediction and mutual-information regularization [16]. INSURE further uses information-theoretic disentanglement and purification to separate class-relevant and auxiliary domain information [17]. Our contribution

is complementary: instead of applying the bottleneck directly to generic deep features, we impose it on physically structured wavelet amplitude subbands.

2.3. Invariant, Domain-Specific, and Causal Representations

Classical domain generalization methods seek invariant feature representations [18–22]. However, recent work shows that domain-specific information can also be useful when it is separated from invariant task information. The mDSDI framework explicitly learns both domain-invariant and domain-specific components to improve generalization [23]. This view is close to our use of Z_{task} and Z_{domain} : domain information is not discarded, but compressed and isolated so that it can support controlled modulation rather than leak into task prediction. The distinction between useful factorization and unsupported statistical independence is also consistent with the broader disentanglement literature, which shows both the utility and the identifiability limitations of unsupervised latent-factor separation [24–26].

2.4. Foundation Models and Test-Time Adaptation

Segmentation foundation models such as SAM [27] and MedSAM [28] improve cross-task transfer by large-scale pretraining and promptability. They address a different axis of generalization from the present work. Our analysis focuses on site and acquisition shifts in a fixed medical segmentation setting, and studies how a physically motivated frequency representation can make such shifts analyzable. Similarly, uncertainty-guided test-time adaptation methods use predictive uncertainty as a deployment-time signal; Section 10 provides a conditional analysis of this idea in the DTCWT amplitude domain.

2.5. Foundational Tools Versus Proposed Contributions

Several equations used below are standard and are included to make the paper self-contained. Table 1 separates established tools from the contributions developed in this work.

Table 1. Relationship between established formulations and the proposed theory.

Component	Status	Role in This paper
Ben-David style DG bounds	Prior work	Motivation for reducing domain divergence
IB/VIB objective	Prior work	Compression language for latent domain codes
Domain cross-entropy loss	Prior work	Supervised identifiability signal for Z_{domain}
Gaussian KL term	Prior work	Computable capacity control for Z_{domain}
FiLM modulation	Prior work	Lightweight conditional modulation operator
DTCWT subband bottleneck	Proposed	Applies VIB to localized amplitude representations
Triple-constraint analysis	Proposed	Identifies distinct failure modes addressed by $\mathcal{L}_{\text{dom}}$, $\mathcal{L}_{\text{KL}}$, and $\mathcal{L}_{\text{orth}}$
Two-pass DTCWT adaptation bound	Proposed	Provides a sufficient condition for uncertainty-guided amplitude correction

To further clarify the methodological position of this paper, Table 2 compares the proposed framework with representative domain generalization and segmentation paradigms. The comparison is not intended to claim state-of-the-art empirical performance; rather, it identifies which mathematical object each framework controls and where the present analysis adds a distinct signal-structured bottleneck.

Table 3 makes the boundary between inherited tools and original theoretical claims explicit. Standard equations are used only as preliminaries or implementation losses; the original contribution is the conditional placement of these tools in a DTCWT-structured phase–amplitude representation.

Table 2. Systematic comparison with representative domain generalization and segmentation frameworks.

Framework	Mechanism and Controlled Quantity	Relation to This Work
Global Fourier DG [3,4]	Global amplitude manipulation controls image-level style.	Motivates frequency-domain DG, but lacks spatial locality; DTCWT uses localized complex subbands.
Adversarial DG [20,21]	Min–max domain confusion controls domain predictability.	Suppresses domain information; our framework retains a compressed domain code Z_{domain} for modulation.
IRM/invariant features [18,19]	Stable predictors control environment-dependent risk.	Does not specify a physical signal decomposition; our premise is tied to medical phase–amplitude structure.
IIB [16]	Mutual-information regularization controls invariant capacity.	Shares the IB language, but applies it to generic features rather than DTCWT amplitude subbands.
mDSDI/INSURE [17,23]	Disentanglement controls invariant/domain-specific factors.	Closest in representation objective; our analysis is conditional on a physically motivated wavelet domain.
SAM/MedSAM [27,28]	Pretraining and prompting control cross-task transfer.	Addresses another generalization axis; our theory targets site/acquisition shifts in a fixed task.
Proposed framework	DTCWT bottleneck controls local domain information, leakage, and calibrated adaptation.	Provides conditional bounds and failure-mode analysis for a signal-structured domain representation.

Table 3. Originality boundary and assumptions of the main theoretical results.

Result	Inherited Component	Proposed Step	Main Assumption or Limitation
Property 2	Classical phase importance and MRI acquisition physics	Formulates a local phase–magnitude complementarity premise for DG analysis	Structural premise; not a universal law and must be empirically checked.
Theorem 1	Data processing and wavelet locality arguments	Compares local DTCWT amplitude information with global FFT-style representations	Requires localized domain shifts and regularity of the domain transform.
Theorem 2	IB/VIB and PAC-Bayesian upper-bound machinery	Applies capacity control to a DTCWT amplitude-domain code Z_{domain}	Bound is loose for deep networks and functions as a capacity-control rationale.
Theorem 5	Linear covariance control	Replaces the stronger independence claim with linear cross-subspace leakage control	Independence follows only under additional linear-Gaussian and sufficiency assumptions.
Theorem 6	Standard domain loss, KL loss, and orthogonality regularization	Identifies the distinct failure mode removed by each constraint	Conditional non-redundancy, not an unconditional disentanglement guarantee.
Theorem 9	Calibration-based uncertainty reasoning	Derives a sufficient condition for uncertainty-guided DTCWT amplitude correction	Depends on calibrated uncertainty transfer under target-domain shift.

3. Preliminaries

3.1. Notation

Let $[n] = \{1, \dots, n\}$. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b}$ denotes the inner product, and $\|\mathbf{a}\|_2 = \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}$. The Frobenius norm of a matrix $\mathbf{M}$ is $\|\mathbf{M}\|_F = \sqrt{\langle \mathbf{M}, \mathbf{M} \rangle_F}$ where $\langle \mathbf{A}, \mathbf{B} \rangle_F = \text{tr}(\mathbf{A}^\top \mathbf{B})$. For a complex number $z = a + ib$, $|z| = \sqrt{a^2 + b^2}$ and $\angle z = \text{atan2}(b, a)$.

We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the multivariate normal distribution and $\text{KL}(\cdot \parallel \cdot)$ for the Kullback–Leibler divergence. The mutual information between random variables X and Y is denoted $I(X; Y)$, following standard information-theoretic notation [29].

3.2. Information Bottleneck Foundations

The equations in this subsection are prior-work foundations included for self-containment. The classical information bottleneck (IB) principle [11] seeks a compressed representation Z of input X that preserves maximal information about a target Y :

$$\min_{p(z|x)} I(X; Z) - \beta I(Z; Y) \quad (9)$$

where $\beta > 0$ trades off compression against prediction.

The variational information bottleneck (VIB) [12] provides a tractable variational approximation:

$$\mathcal{L}_{\text{VIB}} = \mathbb{E}_{x,y} \left[\mathbb{E}_{z \sim q_\phi(z|x)} [-\log p_\theta(y|z)] \right] + \beta \cdot \text{KL}(q_\phi(z|x) \parallel p(z)) \quad (10)$$

where $q_\phi(z|x) = \mathcal{N}(\boldsymbol{\mu}_\phi(x), \boldsymbol{\Sigma}_\phi(x))$ is a learned stochastic encoder and $p(z) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a fixed prior.

3.3. Domain Generalization Bounds

A central result in domain generalization theory is the bound:

$$\forall h \in \mathcal{H} : R_T(h) \leq \frac{1}{K} \sum_{k=1}^K R_k(h) + d(\{P_k\}_{k=1}^K, P_T) + \lambda \quad (11)$$

where $d(\cdot, \cdot)$ measures the divergence between the mixture of source distributions and the target distribution [30].

4. DTCWT: Structured Phase-Amplitude Representation

4.1. Axial-Plane Complex Wavelet Decomposition

For a 3D medical image $\mathbf{x} \in \mathbb{R}^{4 \times D \times H \times W}$ (four-channel MRI with depth D , height H , width W), we define the axial-plane DTCWT slice-wise.

Definition 1 (Axial-Plane 2D DTCWT). For axial slice $z \in [D]$, let $\mathbf{x}_z \in \mathbb{R}^{4 \times H \times W}$. The DTCWT uses a pair of Q -shift dual-tree filter banks $\{h_0^{(a)}, h_1^{(a)}\}$ and $\{h_0^{(b)}, h_1^{(b)}\}$ that form the real and imaginary parts of an approximately analytic complex wavelet. At scale $j \in [J]$ and orientation $\theta \in \{\pm 15^\circ, \pm 45^\circ, \pm 75^\circ\}$, the complex subband coefficient is

$$\mathbf{W}_{j,\theta}(\mathbf{x}_z) = \mathbf{W}_{j,\theta}^{(a)}(\mathbf{x}_z) + i \cdot \mathbf{W}_{j,\theta}^{(b)}(\mathbf{x}_z) \in \mathbb{C}^{4 \times D_j \times H_j} \quad (12)$$

where $D_j = \lfloor D/2^j \rfloor$, $H_j = \lfloor H/2^j \rfloor$ are the downsampled spatial dimensions.

Definition 2 (Subband Amplitude and Phase). For each complex subband $\mathbf{W}_{j,\theta}$, the amplitude and phase are

$$\mathbf{A}_{j,\theta} = |\mathbf{W}_{j,\theta}| \in \mathbb{R}_+^{4 \times D_j \times H_j}, \quad \boldsymbol{\Phi}_{j,\theta} = \angle \mathbf{W}_{j,\theta} \in [-\pi, \pi)^{4 \times D_j \times H_j} \quad (13)$$

Stacking over all slices along the z -axis yields 3D amplitude and phase volumes:

$$\mathbf{A} = \{\mathbf{A}_{j,\theta}\}_{j,\theta} \in \mathbb{R}_+^{\text{Camp} \times D_J \times H_J \times D}, \quad \boldsymbol{\Phi} = \{\boldsymbol{\Phi}_{j,\theta}\}_{j,\theta} \in [-\pi, \pi)^{\text{Camp} \times D_J \times H_J \times D} \quad (14)$$

where $C_{\text{amp}} = 4 \times J \times 6$ is the total number of amplitude channels across all scales and orientations.

Remark 1 (Axial Approximation). *The axial-plane 2D DTCWT processes each slice independently, without explicit through-plane directional selectivity. This is a clinically motivated approximation: MRI is acquired as an axial slice-stack where through-plane resolution is typically coarser than in-plane resolution [31].*

Proposition 2 (Information Loss of Axial Approximation). *Let $\mathcal{W}_{3D}$ denote a true 3D DTCWT with 28 directional subbands and $\mathcal{W}_{2D}^z$ the axial-plane 2D approximation. Under the assumption that through-plane spatial frequencies are bandlimited by the acquisition point-spread function with cutoff ω_z^{max} , the mutual information gap satisfies*

$$I(\mathbf{x}; \mathcal{W}_{3D}(\mathbf{x})) - I(\mathbf{x}; \mathcal{W}_{2D}^z(\mathbf{x})) \leq H\left(\frac{\omega_z^{\text{max}}}{\omega_{xy}^{\text{max}}}\right) \cdot \log(28/6) \quad (15)$$

where $H(p) = -p \log p - (1-p) \log(1-p)$ is the binary entropy function and ω_{xy}^{max} is the in-plane frequency cutoff.

Proof. The gap arises from the 28 vs. 6 directional subbands in the through-plane direction. Under bandlimitedness, the through-plane frequency content contributes proportionally $\omega_z^{\text{max}}/\omega_{xy}^{\text{max}}$ to the total signal entropy. The subband count ratio 28/6 multiplies the directional information capacity. Applying the data processing inequality and the entropy bound yields the result. $\square$

4.2. Phase-Amplitude Coupling and the Need for $\mathcal{L}_{\text{phase}}$

An important technical nuance: DTCWT amplitude and phase are not perfectly independent.

Lemma 1 (Amplitude-Phase Coupling). *Let $\tilde{\mathbf{A}} \neq \mathbf{A}$ be a modified amplitude spectrum. The phase of the reconstructed and re-transformed signal satisfies*

$$\angle \mathcal{W}\left(\mathcal{W}^{-1}(\tilde{\mathbf{A}} \odot e^{i\Phi})\right)_{j,\theta} = \Phi_{j,\theta} + \epsilon_{j,\theta}(\tilde{\mathbf{A}}, \mathbf{A}) \quad (16)$$

where $\epsilon_{j,\theta} \in [-\delta, \delta]^{4 \times D_j \times H_j}$ is a bounded phase drift with $\delta = O(\|\tilde{\mathbf{A}} - \mathbf{A}\|_F / \|\mathbf{A}\|_F)$.

Proof. The DTCWT inverse $\mathcal{W}^{-1}$ involves linear combinations of subband coefficients through synthesis filters. When $\tilde{\mathbf{A}} \neq \mathbf{A}$, the synthesis integration over orientations mixes phase contributions. However, the Q-shift design ensures near-orthogonality between orientations [5], bounding the cross-orientation leakage. Taylor expansion of the phase perturbation yields $\epsilon_{j,\theta} = \nabla_{\mathbf{A}} \Phi_{j,\theta} \cdot (\tilde{\mathbf{A}} - \mathbf{A}) + O(\|\tilde{\mathbf{A}} - \mathbf{A}\|^2)$, where $\|\nabla_{\mathbf{A}} \Phi\| \leq \kappa / \|\mathbf{A}\|$ for some constant κ determined by the filter bank design. $\square$

4.3. Domain Information Concentration in Amplitude Subbands

We now establish the information-theoretic justification for operating the VIB on amplitude subbands.

Theorem 1 (Domain Information in Amplitude Subbands). *Let $D \in [K]$ denote the domain identity random variable and $\mathbf{A}$ the DTCWT amplitude representation. Under Property 2 and mild regularity conditions on $\mathcal{T}_{k \rightarrow l}$, the domain-relevant mutual information satisfies*

$$I(\mathbf{A}; D) \geq I(\mathcal{F}_{\text{FFT}}(\mathbf{x}); D) - \Delta_{\text{FFT}} \quad (17)$$

where $\Delta_{\text{FFT}} = I(\mathbf{x}; D \mid \mathcal{F}_{\text{FFT}}(\mathbf{x}))$ is the domain information lost by the global FFT representation.

Proof Sketch. The global FFT discards spatial locality: $|\mathcal{F}_{\text{FFT}}(\mathbf{x})|(u, v)$ aggregates energy over the entire spatial support. For a spatially localized domain shift (e.g., scanner-dependent bias field in the frontal region), the FFT amplitude at frequency (u, v) receives contributions from both the shifted and unshifted regions, diluting the domain signal. The DTCWT, by contrast, provides spatial localization through its compact-support basis functions. Specifically, the effective spatial support of $\mathbf{W}_{j,\theta}$ at scale j is $O(2^j \times 2^j)$ —much smaller than the full image. By the data processing inequality applied to the spatially localized subbands:

$$I(\mathbf{A}; D) = I(\{\mathbf{A}_{j,\theta}\}_{j,\theta}; D) \geq \max_{j,\theta} I(\mathbf{A}_{j,\theta}; D) = I(D; D) - \Delta_{\text{FFT}} \quad (18)$$

where the inequality follows from the fact that each subband independently captures a portion of the domain information. $\square$

5. Variational Information Bottleneck on Structured Subbands

5.1. Encoder Architecture

We now construct the VIB encoder operating on DTCWT amplitude subbands.

Definition 3 (Amplitude VIB Encoder). Given amplitude volume $\mathbf{A} \in \mathbb{R}_+^{C_{\text{amp}} \times D_J \times H_J \times D}$, the encoder $q_\phi(Z_{\text{domain}}, Z_{\text{task}} \mid \mathbf{A})$ is defined through

$$\mathbf{h} = \text{MLP}_{\text{enc}}(\text{GAP}(\mathbf{A})) \in \mathbb{R}^{d_h} \quad (19)$$

$$\boldsymbol{\mu}, \log \sigma = \text{Linear}_\mu(\mathbf{h}), \text{Linear}_\sigma(\mathbf{h}) \in \mathbb{R}^{d_z} \quad (20)$$

$$Z_{\text{domain}} \sim \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\sigma^2)) \quad (21)$$

$$Z_{\text{task}} = \text{Linear}_{\text{task}}(\mathbf{h}) \in \mathbb{R}^{d_z} \quad (22)$$

where GAP denotes global average pooling over the spatial dimensions and d_z is the latent dimension.

5.2. Variational Bound on Domain Information

The following analysis combines the VIB variational bound with standard information-theoretic and PAC-Bayesian generalization arguments [32,33].

Theorem 2 (VIB Upper Bound on Domain-Attribute Mutual Information). Let $q_\phi(Z_{\text{domain}} \mid \mathbf{A})$ be the encoder defined above with prior $p(Z_{\text{domain}}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. For any test function family $\mathcal{G}$, the mutual information between the domain representation and amplitude satisfies

$$I(Z_{\text{domain}}; \mathbf{A}) \leq \frac{1}{N} \sum_{i=1}^N \text{KL}(q_\phi(Z_{\text{domain}} \mid \mathbf{A}_i) \parallel p(Z_{\text{domain}})) + \log \frac{|\mathcal{G}|}{\delta} \quad (23)$$

with a probability of at least $1 - \delta$ over the training sample of size N , where $|\mathcal{G}|$ is a complexity measure of the encoder class.

Proof. The mutual information admits the variational upper bound

$$\begin{aligned} I(Z_{\text{domain}}; \mathbf{A}) &= \mathbb{E}_{p(\mathbf{A}, Z_{\text{domain}})} \left[\log \frac{q_\phi(Z_{\text{domain}} \mid \mathbf{A})}{p(Z_{\text{domain}})} \right] - \text{KL}(p(Z_{\text{domain}}) \parallel p(Z_{\text{domain}})) \\ &\leq \mathbb{E}_{p(\mathbf{A})} [\text{KL}(q_\phi(Z_{\text{domain}} \mid \mathbf{A}) \parallel p(Z_{\text{domain}}))] \end{aligned} \quad (24)$$

The empirical estimation introduces a finite-sample penalty. By McAllester's PAC-Bayesian bound [32], applied to the KL estimator with function class $\mathcal{G}$ and confidence δ , we obtain the stated result. $\square$

Theorem 2 provides the theoretical justification for using the KL divergence as a computable upper bound on domain-relevant information in the latent representation.

5.3. Information Sufficiency for Segmentation

We must also ensure that compressing domain information does not destroy task-relevant information.

Theorem 3 (Conditional Task Information Preservation). *Let Z_{task} be the deterministic task representation, Z_{domain} the domain representation, and $\mathbf{y}$ the segmentation label. Assume that the combined representation $(Z_{\text{task}}, Z_{\text{domain}})$ is ϵ_s -sufficient for the amplitude representation in the sense that*

$$I(\mathbf{A}; \mathbf{y} \mid Z_{\text{task}}, Z_{\text{domain}}) \leq \epsilon_s, \quad (25)$$

and that residual domain-to-label leakage after conditioning on Z_{task} is bounded by

$$I(Z_{\text{domain}}; \mathbf{y} \mid Z_{\text{task}}) \leq \eta. \quad (26)$$

Then the task representation preserves segmentation-relevant information up to the sufficiency and leakage errors:

$$I(Z_{\text{task}}; \mathbf{y}) \geq I(\mathbf{A}; \mathbf{y}) - \epsilon_s - \eta. \quad (27)$$

Proof. Because $(Z_{\text{task}}, Z_{\text{domain}})$ is generated from $\mathbf{A}$, the chain rule gives

$$I(\mathbf{A}; \mathbf{y}) = I(Z_{\text{domain}}, Z_{\text{task}}; \mathbf{y}) + I(\mathbf{A}; \mathbf{y} \mid Z_{\text{domain}}, Z_{\text{task}}) \quad (28)$$

and

$$I(Z_{\text{domain}}, Z_{\text{task}}; \mathbf{y}) = I(Z_{\text{task}}; \mathbf{y}) + I(Z_{\text{domain}}; \mathbf{y} \mid Z_{\text{task}}). \quad (29)$$

Substituting the second identity into the first and rearranging yields

$$I(Z_{\text{task}}; \mathbf{y}) = I(\mathbf{A}; \mathbf{y}) - I(\mathbf{A}; \mathbf{y} \mid Z_{\text{task}}, Z_{\text{domain}}) - I(Z_{\text{domain}}; \mathbf{y} \mid Z_{\text{task}}). \quad (30)$$

Applying the assumed bounds $I(\mathbf{A}; \mathbf{y} \mid Z_{\text{task}}, Z_{\text{domain}}) \leq \epsilon_s$ and $I(Z_{\text{domain}}; \mathbf{y} \mid Z_{\text{task}}) \leq \eta$ gives the result. This theorem does not require orthogonality to imply statistical independence; orthogonality is used later only as a tractable finite-batch regularizer for reducing linear leakage. $\square$

6. Triple-Constraint Disentanglement Analysis

6.1. Domain Supervision

Definition 4 (Domain Classification Loss). *This is the standard cross-entropy loss for supervised domain classification, used here as an identifiability signal for Z_{domain} rather than as a new loss formulation. Given training domains indexed by $\{1, \dots, K\}$, the domain supervision loss is*

$$\mathcal{L}_{\text{dom}} = -\frac{1}{B} \sum_{b=1}^B \sum_{k=1}^K \mathbf{1}_{[\text{site}(b)=k]} \log \hat{y}_{b,k} \quad (31)$$

where $\hat{\mathbf{y}}_b = \text{softmax}(\text{MLP}_{\text{dom}}(Z_{\text{domain}}^{(b)})) \in \Delta^{K-1}$.

Proposition 3 (Domain Supervision Identifiability). *If $\mathbb{E}[\mathcal{L}_{\text{dom}}] \rightarrow 0$, then as $B \rightarrow \infty$, the posterior $P(D = k \mid Z_{\text{domain}})$ converges to a point mass on the true domain index.*

However, domain supervision alone is *insufficient* for disentanglement.

Proposition 4 (Failure Mode of Pure Domain Supervision). *Let $\mathcal{L}_{\text{dom}}$ be the only domain-related loss term. Then there exist encoder configurations where*

$$I(Z_{\text{domain}}; \mathbf{y}) > 0 \quad \text{and} \quad I(Z_{\text{domain}}; \mathbf{A}) \rightarrow \infty \quad (32)$$

i.e., Z_{domain} captures both domain and task information without bound.

Proof. Construct a pathological encoder that encodes the entire input into Z_{domain} with high precision: $Z_{\text{domain}} = \text{MLP}(\mathbf{A})$ with $\dim(Z_{\text{domain}}) \gg \dim(\mathbf{A})$. Then $\hat{\mathbf{y}}_b$ achieves perfect domain classification while Z_{domain} retains the full input information, including $\mathbf{y}$. The cross-entropy loss provides no penalty for this information leakage. $\square$

6.2. KL Compression

Definition 5 (KL Compression Loss). *This is the standard closed-form Gaussian KL term used in variational bottleneck models. The KL divergence between the domain posterior and the standard normal prior is*

$$\mathcal{L}_{\text{KL}} = \text{KL}(q_{\phi}(Z_{\text{domain}} \mid \mathbf{A}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})) = \frac{1}{2} \sum_{j=1}^{d_z} (\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1) \quad (33)$$

Proposition 5 (KL Prevents Information Overload). *Under the KL constraint, $\mathbb{E}[\mathcal{L}_{\text{KL}}] \leq C$ implies $I(Z_{\text{domain}}; \mathbf{A}) \leq C + \log N$ (by Theorem 2). This prevents the pathological encoder of Proposition 4.*

The $\mathcal{L}_{\text{dom}} + \mathcal{L}_{\text{KL}}$ pair forms a push–pull equilibrium:

- $\mathcal{L}_{\text{dom}}$ pulls Z_{domain} to contain domain-discriminative information;
- $\mathcal{L}_{\text{KL}}$ pushes Z_{domain} toward the uninformative prior, limiting total information capacity.

Theorem 4 (Idealized Push–Pull Equilibrium). *Let $\mathcal{L}_{\text{push-pull}} = \mathcal{L}_{\text{dom}} + \beta \mathcal{L}_{\text{KL}}$. In the idealized constrained information-bottleneck problem, for each attainable capacity level $C(\beta^*)$, there exists a Lagrange multiplier β^* such that any optimal encoder q_{ϕ}^* satisfies*

$$I(Z_{\text{domain}}; D) = \max_{q: I(Z_{\text{domain}}; \mathbf{A}) \leq C(\beta^*)} I(Z_{\text{domain}}; D) \quad (34)$$

where $C(\beta) \rightarrow 0$ as $\beta \rightarrow \infty$ and $C(\beta) \rightarrow \infty$ as $\beta \rightarrow 0$.

Proof. This follows from the constrained optimization perspective of the information bottleneck under the standard regularity conditions for Lagrangian relaxation. The Lagrangian formulation $\min_q I(Z_{\text{domain}}; \mathbf{A}) - \beta I(Z_{\text{domain}}; D)$ is equivalent to $\max_q I(Z_{\text{domain}}; D)$ subject to $I(Z_{\text{domain}}; \mathbf{A}) \leq C(\beta)$. $\mathcal{L}_{\text{dom}}$ approximates $-I(Z_{\text{domain}}; D)$ and $\mathcal{L}_{\text{KL}}$ bounds $I(Z_{\text{domain}}; \mathbf{A})$ (Theorem 2). For neural-network implementations this result should be interpreted as an objective-level rationale rather than a guarantee of global optimization uniqueness. $\square$

However, the push–pull pair still does not guarantee separation between task and domain information within Z_{domain} .

Proposition 6 (Failure Mode of Push–Pull Without Orthogonality). *Even with optimal β^* , there exist encoder configurations where*

$$\langle Z_{\text{task}}, Z_{\text{domain}} \rangle \neq 0 \quad (35)$$

i.e., the task and domain subspaces are not orthogonal, allowing task information to leak into the domain code.

6.3. Orthogonality Constraint

Definition 6 (Orthogonality Loss). *The orthogonality constraint measures the cosine similarity between task and domain representations:*

$$\mathcal{L}_{\text{orth}} = \frac{\|Z_{\text{task}}^\top Z_{\text{domain}}\|_F^2}{\|Z_{\text{task}}\|_F^2 \cdot \|Z_{\text{domain}}\|_F^2} \quad (36)$$

where $Z_{\text{task}}, Z_{\text{domain}} \in \mathbb{R}^{B \times d_z}$ are batch matrices of task and domain latent codes.

Remark 2 (Normalization). Equation (36) uses L_2 -normalized cosine similarity $\in [0, 1]$, decoupling the constraint from batch size. The Frobenius norm $\|Z_{\text{task}}^\top Z_{\text{domain}}\|_F^2$ without normalization would scale linearly with B , introducing unwanted batch-size dependence.

Theorem 5 (Orthogonality Controls Linear Cross-Subspace Leakage). *Let $\tilde{Z}_{\text{task}}$ and $\tilde{Z}_{\text{domain}}$ denote batch-centered task and domain latent matrices. If $\mathcal{L}_{\text{orth}} = 0$, then the empirical cross-covariance between the two latent subspaces vanishes:*

$$\widehat{\text{Cov}}(Z_{\text{task}}, Z_{\text{domain}}) = \frac{1}{B-1} \tilde{Z}_{\text{task}}^\top \tilde{Z}_{\text{domain}} = \mathbf{0}. \quad (37)$$

Consequently, the domain code and task code have no linear cross-subspace leakage within the batch. If, in addition, the latent variables are jointly Gaussian and Z_{domain} is a minimal sufficient statistic for domain identity under the IB objective, this zero cross-covariance implies conditional information separation, $I(Z_{\text{task}}; D \mid \mathbf{y}) = 0$.

Proof. By definition, $\mathcal{L}_{\text{orth}} = 0$ implies $Z_{\text{task}}^\top Z_{\text{domain}} = \mathbf{0}_{d_z \times d_z}$ after normalization. For centered batches this is exactly the empirical cross-covariance up to the factor $(B-1)^{-1}$. Thus no linear predictor can exploit a component of Z_{task} that is aligned with the domain subspace represented by Z_{domain} . This statement does not, by itself, imply statistical independence. The mutual-information statement follows only under the additional linear-Gaussian condition, where zero covariance is equivalent to independence, together with the minimal sufficiency assumption that all domain-discriminative information available to the encoder is allocated to Z_{domain} . $\square$

6.4. Non-Redundancy of the Triple Guarantee

Theorem 6 (Conditional Non-Redundancy of the Triple Constraint). *The three constraints $(\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{KL}}, \mathcal{L}_{\text{orth}})$ address distinct and non-redundant failure modes of the conditional disentanglement objective:*

$$I(Z_{\text{domain}}; D) > 0 \wedge I(Z_{\text{domain}}; \mathbf{A}) \leq \epsilon \wedge \widehat{\text{Cov}}(Z_{\text{task}}, Z_{\text{domain}}) = \mathbf{0}. \quad (38)$$

Under the additional linear-Gaussian and minimal sufficiency assumptions of Theorem 5, the third condition yields conditional information separation.

Proof Structure. We establish the result by identifying the failure mode removed by each constraint:

- (i) Joint effect: $\mathcal{L}_{\text{dom}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{orth}} \rightarrow 0$ implies domain identifiability, bounded latent capacity, and vanishing linear cross-subspace leakage.
- (ii) Necessity of $\mathcal{L}_{\text{dom}}$: Without $\mathcal{L}_{\text{dom}}$, Z_{domain} may collapse to prior $\mathcal{N}(0, I)$, losing all domain information.
- (iii) Necessity of $\mathcal{L}_{\text{KL}}$: Without $\mathcal{L}_{\text{KL}}$, Z_{domain} may encode the full input (Proposition 4).
- (iv) Necessity of $\mathcal{L}_{\text{orth}}$: Without $\mathcal{L}_{\text{orth}}$, domain and task subspaces may overlap (Proposition 6).

Each counterexample is constructed explicitly: the proofs of Propositions 4 and 6 serve as the necessity arguments. $\square$

Theorem 6 is the central theoretical result of this paper. It establishes that each constraint addresses a distinct failure mode. The result is conditional: orthogonality controls linear leakage in general and implies statistical separation only under the additional distributional assumptions stated in Theorem 5.

6.5. Annealing Schedule for β

Definition 7 (KL Annealing). *The coefficient $\beta(t)$ is not a fixed hyperparameter but an annealing schedule:*

$$\beta(t) = \min\left(1, \frac{t}{T_{\text{anneal}}}\right), \quad t \in \{0, \dots, T-1\} \quad (39)$$

This prevents posterior collapse in early training: at $t = 0$, $\beta = 0$ allows the encoder to freely learn informative representations; as $\beta \rightarrow 1$, the KL constraint gradually tightens.

7. Predictive Feature Modulation: $O(1)$ Complexity

7.1. Generative Decoder Infeasibility

In the classical VIB framework, a generative decoder $p_{\theta}(\mathbf{x} | Z)$ reconstructs the input to verify information retention. For our setting:

Proposition 7 (Generative Decoder Complexity). *A generative decoder mapping $Z_{\text{domain}} \rightarrow \mathbf{A}$ requires*

$$\text{params}(\text{Decoder}) = \Omega\left(\prod_{s=1}^3 \dim(\mathbf{A})_s\right) = \Omega(C_{\text{amp}} \times D_J \times H_J \times D) \quad (40)$$

For a 240^3 input with $J = 2$ DTCWT levels, this implies $\dim(\mathbf{A}) \approx 4 \times 2 \times 6 \times 60 \times 60 \times 240 \approx 4.1 \times 10^7$ dimensions, requiring a decoder with $\sim 10^9$ parameters.

7.2. FiLM as Information-Consistent Alternative

Definition 8 (Predictive FiLM Modulation). *The affine modulation form follows the standard FiLM operator; the proposed component is its use as a predictive, bottleneck-conditioned alternative to reconstructing the full DTCWT amplitude volume. Given bottleneck features $\mathbf{F} \in \mathbb{R}^{C \times D_b \times H_b \times W_b}$, the modulated features are*

$$\mathbf{F}_{\text{mod}} = \gamma(Z_{\text{task}}, Z_{\text{domain}}) \odot \mathbf{F} + \beta(Z_{\text{task}}, Z_{\text{domain}}) \quad (41)$$

where $\gamma, \beta \in \mathbb{R}^C$ are channel-wise scale and shift parameters generated by a lightweight hypernetwork $\text{MLP}_{\text{FiLM}}: \mathbb{R}^{2d_z} \rightarrow \mathbb{R}^{2C}$.

Theorem 7 (FiLM Preserves IB Semantics). *The FiLM-based conditional distribution $p(\mathbf{F}_{\text{mod}} | Z_{\text{task}}, Z_{\text{domain}})$ preserves the information bottleneck constraints:*

$$I(Z_{\text{domain}}; \mathbf{F}_{\text{mod}}) \leq I(Z_{\text{domain}}; \mathbf{A}) \leq C(\beta) \quad (42)$$

$$I(Z_{\text{task}}; \mathbf{F}_{\text{mod}}) \geq I(Z_{\text{task}}; \mathbf{y}) - \epsilon \quad (43)$$

where $C(\beta)$ is the KL-imposed capacity bound and ϵ accounts for the FiLM information loss (bounded by the network capacity).

Proof. Equation (42) follows from the data processing inequality: $Z_{\text{domain}} \rightarrow \mathbf{A} \rightarrow \mathbf{F}_{\text{mod}}$ forms a Markov chain. Equation (43): $Z_{\text{task}} \rightarrow \mathbf{F}_{\text{mod}} \rightarrow \mathbf{y}$ is the inference path. The FiLM transformation is Lipschitz-continuous (bounded weights), ensuring $I(Z_{\text{task}}; \mathbf{F}_{\text{mod}}) \geq I(Z_{\text{task}}; \mathbf{y}) - \text{Lip}(\text{FiLM}) \cdot \epsilon$. $\square$

Theorem 8 (Complexity Reduction). *Predictive FiLM modulation achieves*

$$\text{params}(\text{FiLM}) = O(d_z \cdot C) = O(1) \quad (44)$$

compared to $O(N)$ for generative decoding, where $N = \dim(\mathbf{A})$. The spatial complexity reduction factor is:

$$\frac{\text{params}(\text{Decoder})}{\text{params}(\text{FiLM})} = \Omega\left(\frac{\dim(\mathbf{A})}{d_z \cdot C}\right) \quad (45)$$

which is $\sim 10^5$ for typical medical volumes.

Proof. MLP_{FiLM} maps $\mathbb{R}^{2d_z} \rightarrow \mathbb{R}^{2C}$ with hidden dimension d_h , requiring $(2d_z \cdot d_h + d_h \cdot 2C)$ parameters. With $d_z = 128$, $d_h = 64$, $C = 128$: $\text{params} \approx 128 \times 64 + 64 \times 256 \approx 2.5 \times 10^4 = O(1)$. A generative decoder requires $\Omega(\dim(\mathbf{A}))$ parameters to reconstruct the amplitude volume. $\square$

8. Phase Consistency Constraint

Definition 9 (Phase Consistency Loss). *For S subbands total, the phase consistency loss measures cosine similarity between original and post-augmentation DTCWT phases:*

$$\mathcal{L}_{\text{phase}} = 1 - \frac{1}{S} \sum_{s=1}^S \cos(\Phi_s^{\text{orig}}, \Phi_s^{\text{aug}}) \quad (46)$$

where $\cos(\theta_1, \theta_2) = \frac{1}{|\theta_1|} \sum_i \cos(\theta_{1,i} - \theta_{2,i})$ is the mean cosine of angular differences.

Proposition 8 (Phase Consistency as Structural Fidelity). $\mathcal{L}_{\text{phase}} \rightarrow 0$ implies that the post-augmentation reconstruction preserves the DTCWT phase structure:

$$\|\Phi^{\text{orig}} - \Phi^{\text{aug}}\|_{\text{circ}} \rightarrow 0 \quad (47)$$

where $\|\cdot\|_{\text{circ}}$ denotes the circular distance on $[-\pi, \pi)$.

Remark 3 (Domain-Dependence of $\mathcal{L}_{\text{phase}}$). *Unlike Sobel/SSIM-based structure preservation operating in the spatial domain, $\mathcal{L}_{\text{phase}}$ operates within the DTCWT domain, forming a closed optimization loop with the DTCWT front-end. This design ensures that structure preservation is inherently aligned with the operating space of the information bottleneck.*

9. Total Training Objective

Combining all components, the total training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{dis}}(\mathcal{L}_{\text{dom}} + \beta\mathcal{L}_{\text{KL}}) + \alpha\mathcal{L}_{\text{orth}} + \gamma\mathcal{L}_{\text{phase}} \quad (48)$$

Definition 10 (Loss Grouping and Hyperparameter Semantics). *The loss terms are logically grouped into four categories:*

- (i) *Segmentation Bundle (weight 1.0, fixed): $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{seg}}^{\text{Dice}} + \mathcal{L}_{\text{seg}}^{\text{CE}}$;*
- (ii) *Disentanglement Bundle (weight λ_{dis}): $\mathcal{L}_{\text{dom}} + \beta\mathcal{L}_{\text{KL}}$, where β is the annealing coefficient (Equation (39)), not a hyperparameter;*
- (iii) *Orthogonality Term (weight α): $\mathcal{L}_{\text{orth}}$ as defined in Equation (36);*
- (iv) *Phase Term (weight γ): $\mathcal{L}_{\text{phase}}$ as defined above.*

Remark 4 (Effective Hyperparameter Count). *The objective in Equation (48) has three tuneable weights ($\lambda_{\text{dis}}, \alpha, \gamma$), not five, as it may appear. β is the annealing schedule coefficient, and the segmentation weight is fixed at 1.0. This logical grouping reduces the perceived hyperparameter space from $\mathbb{R}_+^5$ to $\mathbb{R}_+^3$.*

10. Two-Pass Test-Time Adaptation Theory

10.1. Uncertainty as Domain Shift Signal

Let $u(\mathbf{x}) \in [0, 1]^{D \times H \times W}$ be the model's predictive uncertainty map (e.g., entropy of softmax outputs). Under domain shift, the model exhibits elevated uncertainty in regions where domain-specific amplitude characteristics mislead the segmentation head. Calibration is evaluated using the standard expected calibration error formulation [34].

Assumption 1 (Calibrated Uncertainty Transfer). *The uncertainty head is ϵ -calibrated in the sense of Expected Calibration Error (ECE):*

$$\text{ECE} = \mathbb{E}_{\hat{p}}[|\mathbb{P}(\hat{y} = y \mid \hat{p} = p) - p|] \leq \epsilon \quad (49)$$

where $\hat{p}$ is the model's predicted confidence.

10.2. Two-Pass Inference Algorithm

The inference procedure first identifies spatial regions with elevated predictive uncertainty, applies a localized correction only to the corresponding DTCWT amplitude coefficients, and then reconstructs the image while preserving its original phase. Algorithm 1 summarizes the complete two-pass procedure used in the subsequent analysis.

Theorem 9 (Two-Pass Generalization Gap Reduction). *Under Assumption 1, the two-pass procedure described in Algorithm 1 achieves*

$$R_T(\text{Two-Pass}) \leq R_T(\text{Single-Pass}) - \Delta_{\text{adapt}} \quad (50)$$

where $\Delta_{\text{adapt}} = \mathbb{E}_{\mathbf{x} \sim P_T}[\mathbb{1}[u_1(\mathbf{x}) > \tau] \cdot \delta_{\text{local}}(\mathbf{x})] \geq 0$, with τ being a confidence threshold and δ_{local} the expected local improvement from amplitude correction in high-uncertainty regions.

Proof Sketch. The proof decomposes the domain generalization error into calibrated and uncalibrated components. Under $\text{ECE} \leq \epsilon$, the uncertainty u_1 serves as a noisy but unbiased signal of domain-induced errors. The soft mask $\mathbf{m}$ localizes corrections to high-uncertainty regions, avoiding unnecessary perturbation of well-generalized regions. The DTCWT amplitude correction $\tilde{\mathbf{A}}$ reduces the local amplitude discrepancy $|\tilde{\mathbf{A}} - \mathbf{A}_T| < |\mathbf{A}_1 - \mathbf{A}_T|$,

where $\mathbf{A}_T$ is the amplitude of a hypothetical in-domain image. Since the model's error increases monotonically with amplitude discrepancy (by the Lipschitz property of the segmentation head), the corrected input yields lower expected error. $\square$

Algorithm 1: Two-Pass Uncertainty-Guided DTCWT Adaptation

Given a trained segmentation model f_θ , an uncertainty head u_θ , and a test image $\mathbf{x}$:

- (1) First pass: compute the initial segmentation logits and uncertainty map,

$$(\hat{\mathbf{y}}_1, u_1) = f_\theta(\mathbf{x}), u_\theta(\mathbf{x}).$$

- (2) Wavelet decomposition: compute DTCWT coefficients

$$\mathcal{W}(\mathbf{x}) = \mathbf{A}_1 \odot e^{i\Phi_1}.$$

- (3) Mask construction: convert the uncertainty map into a smooth spatial mask

$$\mathbf{m} = \text{Smooth}(\mathbf{1}[u_1 > \tau]),$$

where τ is an uncertainty threshold.

- (4) Amplitude correction: apply local amplitude normalization or modulation only inside high-uncertainty regions:

$$\tilde{\mathbf{A}} = (1 - \mathbf{m}) \odot \mathbf{A}_1 + \mathbf{m} \odot \text{Norm}(\mathbf{A}_1).$$

- (5) Reconstruction: reconstruct the adapted image while preserving the original phase:

$$\tilde{\mathbf{x}} = \mathcal{W}^{-1}(\tilde{\mathbf{A}} \odot e^{i\Phi_1}).$$

- (6) Second pass: compute the final prediction

$$\hat{\mathbf{y}}_2 = f_\theta(\tilde{\mathbf{x}}).$$

The output of the procedure is $\hat{\mathbf{y}}_2$.

Theorem 10 (Two-Pass Complexity). *The two-pass procedure requires exactly two forward passes and one DTCWT inverse transform. Compared to Monte Carlo dropout with N samples:*

$$\frac{\text{cost}(\text{Two-Pass})}{\text{cost}(\text{MC Dropout}_N)} = \frac{2}{N} \xrightarrow{N \rightarrow \infty} 0 \quad (51)$$

Proof. Each DTCWT inverse requires $O(D \cdot H \cdot W \cdot \log(HW))$ operations—one FFT per slice. For $240 \times 240 \times 155$ volumes, the cost ratio is $\approx 2/N$. With typical $N = 30$ for MC dropout, the two-pass procedure is $15\times$ faster. $\square$

11. Discussion

11.1. Physical Priors as Dimensionality Relief

Our framework illustrates a general principle: physical priors can reduce the effective dimensionality of representation learning. The DTCWT's structured decomposition effectively sparsifies the information bottleneck problem by pre-separating structure (phase) from style (amplitude). This reduces the effective dimensionality from $O(10^5)$ (raw spatial domain) to $O(10^2)$ (compressed amplitude statistics), enabling the VIB to operate efficiently.

We conjecture that this principle extends beyond medical imaging to any domain where a physically motivated signal decomposition reveals a natural separation of task-relevant and domain-specific information.

11.2. Connections to Domain Generalization Theory

Our work connects to several lines of theoretical DG research:

- (i) Ben-David et al. bound [1]: Under the phase–amplitude premise, the DTCWT front-end can be viewed as a representation mapping g intended to reduce acquisition-driven components of the $\mathcal{H}\Delta\mathcal{H}$ -divergence between source and target domains.
- (ii) Invariant Risk Minimization (IRM) [19]: The triple-constraint mechanism implements a structured invariance objective: task features are encouraged to be invariant across domains, while domain features capture controlled variation.
- (iii) Domain-Adversarial Training [20]: Domain supervision provides an alternative to adversarial training, avoiding the min–max instability while explicitly retaining a compressed domain code.

11.3. Verifiability

A defining feature of our theoretical framework is its verifiability through clear experimental predictions:

- (1) Amplitude-phase swap: If Property 2 is not useful in a given dataset, then amplitude swapping between domains should not yield the predicted acquisition-style perturbations and may substantially degrade segmentation.
- (2) Ablation non-redundancy: If Theorem 6 is reflected in a practical implementation, then progressive ablation is expected to show: $(B) < (A)$, $(C) > (B)$, $(D) > (C)$, $(E) > (D)$ —where violations would indicate that the corresponding constraint is not contributing as predicted.
- (3) VIB compression: If the VIB is effective, the domain classifier accuracy should remain high while KL compression increases over training, indicating that domain information is being compressed rather than discarded.

11.4. Validation Logic

The two empirical checks below are designed to test different parts of the theoretical argument rather than to serve as a complete benchmark. The FeTS site-classification experiment tests the premise that DTCWT amplitude carries more domain/site information than phase. The BraTS segmentation experiment tests a weaker but necessary implementation condition: a DTCWT front end must remain compatible with a real public medical segmentation task and must not collapse segmentation performance. Table 4 summarizes how each check maps to the corresponding theoretical claim.

Table 4. How the illustrative empirical checks map to the theoretical claims.

Theoretical Issue	Empirical Check	Positive Evidence Expected	Remaining Limitation
Amplitude carries site information	FeTS site classification from DTCWT summaries	Amplitude features predict site better than phase features and above chance	Does not prove the premise for every scanner, anatomy, or modality
Phase is less domain-specific	Same FeTS comparison using phase features	Phase features show weaker site prediction than amplitude features	Phase can still carry domain information under misregistration or structural shift
DTCWT front end is usable for segmentation	BraTS DTCWT-UNet validation	Non-trivial Dice on public brain-tumor segmentation cases	Not a controlled SOTA comparison and not a full ablation of all losses
Theoretical claims are conditional	Conservative interpretation of both checks	Results support plausibility rather than universal validity	Broader validation across centers and modalities remains necessary

11.5. Illustrative Empirical Sanity Check

Although the main contribution of this paper is theoretical, we performed a lightweight sanity validation of the phase–amplitude premise using the FeTS 2022 multi-site brain tumor segmentation data [35]. The purpose is not to provide a full segmentation benchmark, but to test whether DTCWT amplitude statistics carry more site/domain information than DTCWT phase statistics. This directly addresses the weakest point of the theory: if amplitude and phase were equally site-discriminative, then the proposed phase–amplitude separation would have little empirical support.

We selected five FeTS sites with sufficient sample counts, sampled eight cases per site, and extracted five non-empty axial slices per case. For each case, we computed a two-level axial DTCWT and summarized amplitude and phase subbands using simple per-channel and per-orientation statistics. A balanced logistic regression classifier was then trained to predict site identity using stratified five-fold cross-validation. Table 5 reports the resulting domain classification performance.

Table 5. Sanity validation of the DTCWT phase–amplitude premise on FeTS 2022 site classification. Chance accuracy is 0.20 because five sites are used. Values are mean $\pm$ standard deviation over stratified five-fold cross-validation.

Feature Set	Accuracy	Macro-F1
DTCWT amplitude statistics	0.700 $\pm$ 0.127	0.668 $\pm$ 0.184
DTCWT phase statistics	0.450 $\pm$ 0.127	0.405 $\pm$ 0.092
Amplitude + phase statistics	0.775 $\pm$ 0.184	0.695 $\pm$ 0.247

These results support the structural premise that site/domain information is more concentrated in DTCWT amplitude features than in phase features. They should be interpreted conservatively: the experiment validates the plausibility of the representation assumption under the sampled FeTS setting, but it does not replace a full segmentation benchmark or prove that the premise holds under all medical imaging shifts. The result is nevertheless informative because it tests the premise using only low-capacity summary features and a simple linear classifier; the observed gap is therefore not a consequence of a large segmentation network learning arbitrary site cues.

11.6. Illustrative Public-Dataset Segmentation Check

We also evaluated whether the DTCWT front-end can be instantiated in a public medical image segmentation setting. Using the public BraTS 2023 adult glioma challenge data [36], we trained DTCWT-UNet variants with one, two, and three DTCWT decomposition levels and evaluated them on the held-out internal validation split used by our implementation. All three completed runs used the same backbone width, training schedule, and parameter budget (17.5 M trainable parameters; 150 epochs; batch size 2). The evaluation included 251 validation cases and reports per-case Dice scores for the BraTS tumor regions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). This experiment addresses a different concern from the FeTS site-classification test: it checks that the DTCWT representation can be used in an actual public segmentation workflow, rather than only in an abstract site-classification analysis. The results of the three completed DTCWT-UNet configurations are reported in Table 6.

This segmentation check is deliberately limited: it verifies that DTCWT-based representations can support non-trivial public brain-tumor segmentation performance, but it is not presented as a controlled state-of-the-art comparison. Only completed and internally consistent DTCWT-front-end runs are reported. Together with Table 5, the results provide

empirical support for the central representation premise while preserving the paper’s main scope as a conditional mathematical analysis.

Table 6. Illustrative BraTS 2023 segmentation check for completed DTCWT-front-end configurations. Values are per-case Dice on 251 validation cases. Mean Dice is reported as mean $\pm$ standard deviation; WT, TC, and ET columns report class-wise means.

Config.	Cases	Mean Dice	WT	TC	ET
DTCWT-UNet, $J = 1$	251	0.717 ± 0.172	0.649	0.750	0.752
DTCWT-UNet, $J = 2$	251	0.731 ± 0.157	0.658	0.782	0.752
DTCWT-UNet, $J = 3$	251	0.725 ± 0.163	0.654	0.770	0.751

11.7. Limitations

We note several theoretical limitations:

- (i) Axial-plane approximation: The 2D axial DTCWT provides an upper bound on through-plane information loss, but the bound’s tightness depends on the acquisition parameters.
- (ii) ECE assumption: Theorem 9 depends on calibrated uncertainty, which may not hold under extreme domain shifts.
- (iii) PAC-Bayesian bounds: The finite-sample bounds (Theorem 2) involve complexity terms that are loose for deep networks, suggesting room for tighter analysis using modern generalization theory.
- (iv) Empirical scope: The FeTS and BraTS checks are illustrative validations of the representation premise and implementation feasibility. They do not constitute a full multi-dataset benchmark, do not establish state-of-the-art segmentation performance, and do not test every component of the full training objective.
- (v) Domain-shift scope: The phase–amplitude premise is most appropriate for acquisition-driven appearance shifts after spatial normalization. It may be weaker for cross-modality shifts, severe pathology-induced deformation, missing sequences, or shifts where anatomy and acquisition style are entangled.
- (vi) Independence claims: The orthogonality constraint controls linear cross-subspace leakage in finite batches. Statistical independence or mutual-information separation requires additional distributional and sufficiency assumptions; the revised theorems state this explicitly.

12. Conclusions

We have presented a mathematical framework for domain generalization motivated by complex wavelet decomposition and variational information theory. The key theoretical contributions include: (1) formalization of the local phase-magnitude complementarity premise that motivates DTCWT as a front-end for domain analysis; (2) a conditional analysis showing when amplitude subbands concentrate domain-relevant information more effectively than global Fourier representations; (3) a triple-constraint non-redundancy analysis showing that domain supervision, KL compression, and orthogonality address distinct failure modes; (4) a predictive modulation scheme achieving $O(1)$ spatial parameter complexity with respect to the amplitude volume while preserving information bottleneck semantics; and (5) a two-pass adaptation theory with a sufficient condition for generalization gap reduction under calibration assumptions.

The framework is designed to be verifiable: its main assumptions and conditional results map to concrete experimental predictions. The illustrative FeTS and BraTS checks do not convert the paper into a complete empirical benchmark, but they test whether the central phase–amplitude premise is plausible and whether a DTCWT front end can

support public medical segmentation. While developed in the context of medical image segmentation, the principles of physically motivated information bottleneck disentanglement may extend to other domains where signal structure can reduce the dimensionality of representation learning.

Author Contributions: Conceptualization, F.L. and Z.W.; methodology, F.L.; formal analysis, F.L.; writing—original draft preparation, F.L.; writing—review and editing, F.L. and Z.W.; supervision, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new datasets were created in this study. The illustrative sanity validation used the publicly available FeTS 2022 data, and the segmentation feasibility check used the public BraTS 2023 challenge data. The code, case lists, experiment provenance notes, and derived summary statistics supporting the illustrative validation checks are available at <https://github.com/frankliuf/entropy-dtcwt-revision-artifacts> (accessed on 28 June 2026). Raw FeTS and BraTS medical images are not redistributed in that repository and should be obtained from their official data-access portals under the corresponding data-use agreements.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DTCWT	Dual-Tree Complex Wavelet Transform
VIB	Variational Information Bottleneck
DG	Domain Generalization
KL	Kullback–Leibler
IB	Information Bottleneck
FiLM	Feature-Wise Linear Modulation
ECE	Expected Calibration Error
FFT	Fast Fourier Transform
IRM	Invariant Risk Minimization
MRI	Magnetic Resonance Imaging

References

1. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175.
2. Oppenheim, A.V.; Lim, J.S. The importance of phase in signals. *Proc. IEEE* **1981**, *69*, 529–541.
3. Yang, Y.; Soatto, S. FDA: Fourier domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
4. Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; Tian, Q. A Fourier-based framework for domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
5. Selesnick, I.W.; Baraniuk, R.G.; Kingsbury, N.G. The dual-tree complex wavelet transform. *IEEE Signal Process. Mag.* **2005**, *22*, 123–151.
6. Kingsbury, N.G. Complex wavelets for shift invariant analysis and filtering of signals. *Appl. Comput. Harmon. Anal.* **2001**, *10*, 234–253.
7. Miller, E.L.; Kingsbury, N.G. The design of multidimensional filters for the dual-tree complex wavelet transform. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Detroit, MI, USA, 9–12 May 1995.
8. Daubechies, I. *Ten Lectures on Wavelets*; SIAM: Philadelphia, PA, USA, 1992.
9. Mallat, S. *A Wavelet Tour of Signal Processing*, 3rd ed.; Academic Press: Cambridge, MA, USA, 2008.

10. Desai, S.; Lakshminarayanan, B.D.S. Domain generalization via frequency-domain data augmentation. *arXiv* **2021**, arXiv:2103.02503.
11. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, arXiv:physics/0004057.
12. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep variational information bottleneck. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
13. Achille, A.; Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2897–2905.
14. Shwartz-Ziv, R.; Tishby, N. Opening the black box of deep neural networks via information. *arXiv* **2017**, arXiv:1703.00810.
15. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the information bottleneck theory of deep learning. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
16. Li, B.; Shen, Y.; Wang, Y.; Zhu, W.; Reed, C.; Li, D.; Keutzer, K.; Zhao, H. Invariant information bottleneck for domain generalization. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 7399–7407.
17. Yu, X.; Tseng, H.-H.; Yoo, S.; Ling, H.; Lin, Y. INSURE: An information theory inspired disentanglement and purification model for domain generalization. *IEEE Trans. Image Process.* **2024**, *33*, 3508–3519.
18. Muandet, K.; Balduzzi, D.; Schölkopf, B. Domain generalization via invariant feature representation. In Proceedings of the International Conference on Machine Learning (ICML), Atlanta, GA, USA, 16–21 June 2013.
19. Arjovsky, M.; Bottou, L.; Gulrajani, I.; Lopez-Paz, D. Invariant risk minimization. *arXiv* **2019**, arXiv:1907.02893.
20. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 2096–2030.
21. Li, H.; Pan, S.J.; Wang, S.; Kot, A.C. Domain generalization with adversarial feature learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
22. Verma, S.; Zhou, Z.-H. Domain generalization with domain-specific aggregation modules. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
23. Bui, M.-H.; Tran, T.; Tran, A.; Phung, D. Exploiting domain-specific features to enhance domain generalization. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021; Volume 34, pp. 21189–21201.
24. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. β -VAE: Learning basic visual concepts with a constrained variational framework. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
25. Chen, T.Q.; Li, X.; Grosse, R.B.; Duvenaud, D.K. Isolating sources of disentanglement in variational autoencoders. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018.
26. Locatello, F.; Bauer, S.; Lucic, M.; Rätsch, G.; Gelly, S.; Schölkopf, B.; Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019.
27. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 4015–4026.
28. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 654.
29. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2006.
30. Albuquerque, I.; Monteiro, J.; Falk, T.H.; Mitliagkas, I. Generalizing to unseen domains via distribution matching. *arXiv* **2020**, arXiv:1911.00804.
31. Brown, R.W.; Cheng, Y.-C.N.; Haacke, E.M.; Thompson, M.R.; Venkatesan, R. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*; Wiley: Hoboken, NJ, USA, 2014.
32. McAllester, D.A. PAC-Bayesian stochastic model selection. *Mach. Learn.* **2003**, *51*, 5–21.
33. Riedel, P.; Hénon, S.F.B. Information-theoretic generalization bounds for deep learning. *arXiv* **2018**, arXiv:1810.04161.
34. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017.
35. Pati, S.; Baid, U.; Edwards, B.; Sheller, M.; Wang, S.-H.; Reina, G.A.; Foley, P.; Gruzdev, A.; Karkada, D.; Davatzikos, C.; et al. Federated learning enables big data for rare cancer boundary detection. *Nat. Commun.* **2022**, *13*, 7346.
36. BraTS 2023 Challenge. Synapse: Syn51156910. Available online: <https://www.synapse.org/Synapse:syn51156910/wiki/622356> (accessed on 23 June 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.