FreqMedCLIP: Frequency Information for Text-Guided Medical Image Segmentation

Ngo Thanh Trung[1]*, Tran Long Vu[1]**, and Doan Hoang Long[1]***

School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
{Trung.NT227990, Vu.TL228054, Long.DH228027}@sis.hust.edu.vn

Abstract. Medical image segmentation requires both reliable boundary detection and semantic understanding of anatomical structures. Vision transformers capture global context well but suppress high-frequency image components through attention smoothing, producing blurred boundaries. Frequency-domain methods preserve edge detail but lack the semantic awareness to tell meaningful boundaries from noise. We propose *FreqMedCLIP*, a dual-stream architecture that couples a frozen Biomed-CLIP vision transformer with a learned frequency encoder operating on Haar wavelet sub-bands. A bidirectional cross-attention module (FFBI) enables symmetric information exchange between streams at a shared bottleneck, while language-guided FiLM conditioning (LFFI) provides text-driven semantic guidance at each decoder resolution. Experiments on brain tumor, breast lesion, and lung nodule segmentation benchmarks show that FreqMedCLIP achieves 0.867 average Dice coefficient (vs. 0.845 for SAM-Med2D and 0.805 for the semantic-only baseline) and 2.1 mm Hausdorff distance (vs. 6.1 mm for SAM-Med2D). Ablations confirm that the frequency encoder contributes +4.2% Dice, FFBI fusion adds +3.0%, and LFFI text guidance adds +2.1%.

Keywords: medical image segmentation · frequency-domain analysis · dual-stream architecture · text-guided segmentation · cross-modal fusion

1 Introduction

Medical image segmentation underpins clinical workflows from tumor delineation to organ-at-risk contouring and lesion monitoring [14]. The task is dual in nature: a model must *recognise* what it segments, establishing semantic correspondence between image regions and anatomical categories, and *locate* the precise spatial extent of those structures, including fine boundary details that carry direct clinical weight. These two demands engage different parts of the image signal. Semantic recognition depends on long-range contextual cues and global intensity

* Equal contribution

** Equal contribution

*** Equal contribution

distributions. Boundary delineation requires sensitivity to local high-frequency transitions at tissue interfaces.

Convolutional architectures, led by U-Net [25] and its descendants [6,21], address this tension through hierarchical feature representations and skip connections. The inductive biases of convolution, however, limit capacity to model the global context needed to distinguish, for example, a necrotic tumor core from surrounding oedema in a glioma scan. Vision transformers (ViTs) replace local convolutions with self-attention over image patches [4], enabling long-range dependency modelling. Hybrid designs such as TransUNet [2] combine both paradigms. Foundation models built on ViT pretraining, including SAM [7] and its medical adaptations [16], as well as BiomedCLIP [28], have shown strong zero-shot and few-shot segmentation. Text-image alignment through CLIP-style training [24], extended by LViT [13] and MedCLIP-SAM [8], lets models condition segmentation on natural-language descriptions of the target structure.

Yet transformer-based architectures inherit a systematic weakness: their patch-level operation and attention smoothing suppress high-frequency image components that delineate anatomical boundaries. Attention maps across deep ViT layers tend toward low spatial frequency, capturing coarse layouts rather than boundary topology [22]. The result is segmentations that correctly identify the gross location of a structure but yield blurred or misaligned boundaries, inflating boundary-sensitive metrics such as Hausdorff distance. Frequency-domain methods have long been used for edge detection. Wavelet decomposition, Fourier analysis, and Laplacian operators all isolate high-frequency content [18]. They lack the semantic context, though, to distinguish meaningful anatomical boundaries from imaging noise or tissue texture.

This paper asks whether a single architecture can achieve strong semantic segmentation *and* reliable boundary localisation by explicitly coupling the complementary strengths of semantic and frequency pathways. We answer yes with *FreqMedCLIP*, a dual-stream architecture that pairs a frozen BiomedCLIP vision encoder with a learned frequency encoder operating on Haar wavelet sub-bands. The two streams interact at a shared bottleneck through bidirectional cross-attention (FFBI), and text embeddings derived from natural-language prompts modulate both decoder branches via FiLM conditioning at each spatial scale. The design rests on a signal-processing principle: low-frequency components encode global semantic structure while high-frequency components encode edges and fine detail. Fusing both under language guidance yields representations that are simultaneously semantically grounded and boundary-aware. On three diverse benchmarks (brain, breast, and lung segmentation), FreqMedCLIP achieves 0.867 average Dice coefficient and 2.1 mm mean Hausdorff distance, compared with 0.845 Dice and 6.1 mm HD for SAM-Med2D.

The principal contributions of this paper are:

1. **A dual-stream frequency-semantic architecture.** We couple a pretrained biomedical vision transformer with a ConvNeXt-based frequency encoder fed on one-level Haar DWT sub-bands. The frequency front-end gives computationally efficient, theoretically grounded access to directional edge

information that the ViT systematically suppresses. On three benchmarks, the dual-stream design improves average Dice by 6.2% over the semantic-only baseline and reduces Hausdorff distance from 8.2 mm to 2.1 mm.

2. **Bidirectional frequency-feature bridging integration (FFBI).** A bottleneck cross-attention module that enables symmetric information exchange between the semantic and frequency streams. Each stream selectively queries the other for complementary information, yielding representations that neither could produce alone. FFBI contributes 3.0% absolute Dice improvement over concatenation-based fusion in ablation studies.
3. **Language-guided feature fusion via FiLM modulation (LFFI).** Text embeddings from the BiomedCLIP language encoder modulate both decoder branches at every spatial resolution ($14 \to 28 \to 56 \to 112$), providing progressive semantic guidance that prevents semantic drift during upsampling. LFFI contributes 2.1% Dice improvement and is most effective at intermediate decoder scales.

2 Related Work

Vision Transformers and Foundation Models for Medical Segmentation. The Vision Transformer (ViT) [4] showed that self-attention over patch sequences can match or exceed convolutional networks for image recognition. Park and Kim [22] found that ViTs learn low-frequency features more readily than CNNs, with attention maps exhibiting strong spatial smoothness, a property beneficial for semantic understanding but harmful for boundary localisation. TransUNet [2] introduced a hybrid CNN-transformer encoder for medical segmentation, establishing the value of combining local and global representations. In the medical domain, foundation models trained on large multimodal corpora have emerged as strong backbones. BiomedCLIP [28] aligns biomedical image and text representations through contrastive pretraining on fifteen million scientific image-text pairs, producing semantically rich features that transfer well to segmentation. SAM [7] and its medical adaptation SAM-Med2D [3] offer strong interactive segmentation but lack inherent frequency-domain awareness. Our approach adopts BiomedCLIP as a frozen semantic backbone and complements it with an explicit frequency pathway, rather than relying on a single foundation model for all signal components.

Frequency-Domain Analysis in Deep Learning. Classical frequency analysis, including Fourier transforms, Laplacian operators, and wavelet decompositions, separates images into sub-bands that isolate different spatial scales and orientations [18]. In deep learning, frequency-domain methods have been applied to domain adaptation [27], network robustness analysis, and medical segmentation. DUWS-Net [12] proposes a wavelet-based dual U-Net that fuses spatial and frequency representations through a spatial-frequency transformer. It shares architectural motivation with our dual-stream design but lacks language guidance. Huang and Zhou [5] address class imbalance through frequency re-weighting, treating frequency as an auxiliary signal rather than an independent encoding pathway. Our approach applies a learned encoder directly to multi-sub-band wavelet inputs, letting the network discover task-optimal frequency representations end-to-end while retaining the interpretability of the DWT decomposition.

Dual-Stream and Multi-Scale Fusion Architectures. Decomposing a task into complementary processing pathways has a long history in medical image analysis. nnU-Net [6] showed that careful single-stream U-Net training can match heavily engineered multi-stream systems, establishing a strong baseline. Attention U-Net [21] introduced gating mechanisms for selective skip-connection activation, anticipating the need for principled rather than unconditional feature fusion. In the transformer era, dual-pathway designs have been revisited: complementary pathways processing different resolutions or modalities exchange information through cross-attention rather than concatenation [13]. The key insight is that bidirectional cross-attention at a shared bottleneck (the mechanism underlying our FFBI module) allows each pathway to selectively query the other for complementary information. This yields representations that neither stream could produce alone, which differs qualitatively from additive or concatenation fusion that lacks the selectivity afforded by attention.

Text-Guided Segmentation. CLIP [24] established text-image alignment at scale, enabling language-conditioned vision systems. LViT [13] integrates text into a U-Net-style architecture via cross-attention at the encoder, showing consistent improvements over visual-only baselines in medical segmentation. MedCLIP-SAM [8] and MedCLIP-SAMv2 [9] bridge CLIP-based text features with SAM's prompt-driven segmentation, achieving strong zero-shot performance on diverse medical tasks. Li et al. [11] show that text prompts can resolve ambiguity in segmenting structures with overlapping appearance, an observation that motivates our use of language guidance at multiple decoder scales. Our LFFI modules apply FiLM conditioning [23] at each spatial resolution of both decoder branches, letting the text signal modulate boundary refinement progressively. This design has not been explored in prior text-guided segmentation work, where language conditioning is typically applied at a single scale or only at the encoder level.

3 Proposed Method

3.1 Overview

We propose a dual-stream text-guided medical image segmentation framework that disentangles (i) *global semantic understanding* and (ii) *frequency-driven boundary cues*, then couples them through a bottleneck interaction module and a dual-decoder design. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ and a text prompt p, the model predicts a binary mask $\mathbf{M} \in [0,1]^{H \times W}$. Fig. 1 provides an architectural overview.

[Architecture diagram placeholder — replace with Arch.png before submission]

Fig. 1: **Architecture overview.** FreqMedCLIP consists of (1) a BiomedCLIP ViT semantic encoder with an FPN adapter (top stream), (2) a Haar DWT + ConvNeXt-Tiny frequency encoder (bottom stream), (3) a bottleneck bidirectional fusion block (FFBI), (4) two symmetric text-conditioned decoders, and (5) a learnable fusion head for final mask prediction.

Pipeline summary. The overall computation proceeds as:

$$\mathbf{S}_{1:4} = \mathrm{FPN}\big(\mathrm{SemEnc}(\mathbf{I}, p)\big), \quad \mathbf{F}_{\mathrm{stem},1:3} = \mathrm{FreqEnc}(\mathrm{DWT}(\mathbf{I})), \quad [\hat{\mathbf{S}}_1, \hat{\mathbf{F}}_3] = \mathrm{FFBI}(\mathbf{S}_1, \mathbf{F}_3), \tag{1}$$

$$\mathbf{L}_s = \mathrm{Dec}_s(\hat{\mathbf{S}}_1, \mathbf{S}_{2:4}, p), \quad \mathbf{L}_f = \mathrm{Dec}_f(\hat{\mathbf{F}}_3, \mathbf{F}_{\mathrm{stem},1:2}, p), \quad \mathbf{M} = \sigma\Big(\mathrm{Fuse}(\mathbf{L}_s, \mathbf{L}_f)\Big). \tag{2}$$

Notation. The prompt p is encoded into token embeddings $\mathbf{T} \in \mathbb{R}^{L \times d_t}$ and a pooled embedding $\bar{\mathbf{t}} \in \mathbb{R}^{d_t}$. For a feature map $\mathbf{X} \in \mathbb{R}^{C \times h \times w}$, $\mathrm{flat}(\mathbf{X}) \in \mathbb{R}^{(hw) \times C}$ denotes spatial flattening into tokens.

3.2 Input Processing

We resize each image to 224×224 for the ViT-based semantic encoder (consistent with CLIP-style training), and normalise intensities per modality (z-score for MRI, min-max for CT). The predicted mask is rescaled to the original resolution using bilinear interpolation.

3.3 Text Encoder

We encode the prompt using the BiomedCLIP text encoder:

$$\mathbf{T} = \mathrm{TextEnc}(\mathrm{tokenize}(p)) \in \mathbb{R}^{L \times d_t}, \qquad \bar{\mathbf{t}} = \mathrm{Pool}(\mathbf{T}) \in \mathbb{R}^{d_t}, \tag{3}$$

where Pool(·) extracts the pooled end token. We use $\mathbf{T}$ for token-level cross-modal interaction and $\bar{\mathbf{t}}$ for lightweight FiLM [23] modulation throughout the decoders.

3.4 Semantic Stream: BiomedCLIP ViT with FPN Adapter

Multi-layer ViT feature extraction. The BiomedCLIP vision transformer processes $\mathbf{I}$ and returns hidden states $\{\mathbf{H}^{(\ell)}\}_{\ell=0}^{12}$. We extract four layers $\ell \in \{12, 9, 6, 3\}$ (deep-to-shallow), remove the class token, and reshape to spatial grids:

$$\mathbf{U}^{(\ell)} = \mathrm{reshape}\Big(\mathbf{H}_{1:}^{(\ell)}\Big) \in \mathbb{R}^{d_v \times 14 \times 14}, \quad d_v = 768. \tag{4}$$

Text-Conditioned Adapters (TCA). To enable early text–vision interaction inside the encoder, we insert lightweight Text-Conditioned Adapters (TCA) after selected ViT layers $\{3, 6, 9, 12\}$. TCA performs token-level cross-attention from image tokens to text tokens:

$$\mathbf{X}_{\mathrm{ca}}^{(\ell)} = \mathbf{X}^{(\ell)} + \mathrm{CA}\big(\mathbf{X}^{(\ell)}\mathbf{W}_Q,\ \mathbf{T}\mathbf{W}_K,\ \mathbf{T}\mathbf{W}_V\big), \tag{5}$$

followed by a small residual FFN:

$$\mathbf{X}_{\mathrm{tca}}^{(\ell)} = \mathbf{X}_{\mathrm{ca}}^{(\ell)} + \mathrm{FFN}\big(\mathrm{LN}(\mathbf{X}_{\mathrm{ca}}^{(\ell)})\big). \tag{6}$$

When full cross-attention is too costly, TCA can be replaced by FiLM-on-tokens: $\mathbf{X}_{\mathrm{tca}}^{(\ell)} = \boldsymbol{\gamma}^{(\ell)} \odot \mathbf{X}^{(\ell)} + \boldsymbol{\beta}^{(\ell)}$, where $[\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{\beta}^{(\ell)}] = \mathrm{MLP}^{(\ell)}(\bar{\mathbf{t}})$.

FPN Adapter: semantic pyramid $\mathbf{S}_1 \to \mathbf{S}_4$*.* Since ViT features are isotropic (14×14), we construct a 4-level pyramid through progressive 2× upsampling with lateral fusion:

$$\begin{aligned}
\mathbf{S}_1 &= \phi_{12}(\mathbf{U}^{(12)}) \in \mathbb{R}^{C_1 \times 14 \times 14}, \\
\mathbf{S}_2 &= \phi_9\Big(\mathrm{Up}_2(\mathbf{S}_1) \oplus \psi_9(\mathbf{U}^{(9)})\Big) \in \mathbb{R}^{C_2 \times 28 \times 28}, \\
\mathbf{S}_3 &= \phi_6\Big(\mathrm{Up}_2(\mathbf{S}_2) \oplus \psi_6(\mathbf{U}^{(6)})\Big) \in \mathbb{R}^{C_3 \times 56 \times 56}, \\
\mathbf{S}_4 &= \phi_3\Big(\mathrm{Up}_2(\mathbf{S}_3) \oplus \psi_3(\mathbf{U}^{(3)})\Big) \in \mathbb{R}^{C_4 \times 112 \times 112}.
\end{aligned} \tag{7}$$

ψ_ℓ are 1×1 channel-alignment projections; ϕ_ℓ are 3×3 refinements. $\mathbf{S}_1$ feeds FFBI; $\mathbf{S}_{2:4}$ serve as skip connections for the semantic decoder.

3.5 Frequency Stream: Haar DWT + ConvNeXt-Tiny

Haar DWT decomposition. We apply a one-level 2D discrete wavelet transform (Haar) to explicitly separate frequency bands:

$$[\mathbf{I}_{LL}, \mathbf{I}_{LH}, \mathbf{I}_{HL}, \mathbf{I}_{HH}] = \mathrm{DWT}_{\mathrm{Haar}}(\mathbf{I}). \tag{8}$$

LL preserves coarse anatomy; $LH/HL/HH$ emphasise directional edges and fine detail. We concatenate sub-bands channel-wise: $\mathbf{I}_{\mathrm{dwt}} = \mathrm{concat}(\mathbf{I}_{LL}, \mathbf{I}_{LH}, \mathbf{I}_{HL}, \mathbf{I}_{HH})$.

ConvNeXt-Tiny frequency encoder pyramid. ConvNeXt-Tiny [15] serves as the frequency encoder backbone (first convolution modified to accept $\mathbf{I}_{\mathrm{dwt}}$):

$$[\mathbf{F}_{\mathrm{stem}}, \mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3] = \mathrm{ConvNeXtTiny}(\mathbf{I}_{\mathrm{dwt}}), \tag{9}$$

where $\mathbf{F}_{\mathrm{stem}} \in \mathbb{R}^{C_s \times 112 \times 112}$, $\mathbf{F}_1 \in \mathbb{R}^{C'_1 \times 56 \times 56}$, $\mathbf{F}_2 \in \mathbb{R}^{C'_2 \times 28 \times 28}$, $\mathbf{F}_3 \in \mathbb{R}^{C'_3 \times 14 \times 14}$. $\mathbf{F}_3$ is fused with $\mathbf{S}_1$ in FFBI; $\mathbf{F}_{\mathrm{stem}}$, $\mathbf{F}_1$, $\mathbf{F}_2$ provide skip features for the frequency decoder.

3.6 FFBI: Bottleneck Fusion Between Semantic and Frequency Streams

The Frequency–Feature Bridging Integration (FFBI) couples the two streams at the 14×14 bottleneck. Both feature maps are flattened into token sequences and projected to a shared width d:

$$\tilde{\mathbf{Z}}_s = \mathrm{flat}(\mathbf{S}_1)\mathbf{W}_s, \quad \tilde{\mathbf{Z}}_f = \mathrm{flat}(\mathbf{F}_3)\mathbf{W}_f, \tag{10}$$

then fused via bidirectional cross-attention:

$$\hat{\mathbf{Z}}_s = \tilde{\mathbf{Z}}_s + \mathrm{CA}\Big(\tilde{\mathbf{Z}}_s, \tilde{\mathbf{Z}}_f, \tilde{\mathbf{Z}}_f\Big), \qquad \hat{\mathbf{Z}}_f = \tilde{\mathbf{Z}}_f + \mathrm{CA}\Big(\tilde{\mathbf{Z}}_f, \tilde{\mathbf{Z}}_s, \tilde{\mathbf{Z}}_s\Big). \tag{11}$$

This allows the semantic stream to query boundary cues from the frequency stream and vice versa, without text bias at the fusion point. Optionally, a text gate is applied post-fusion:

$$[\boldsymbol{\gamma}_b, \boldsymbol{\beta}_b] = \mathrm{MLP}_b(\bar{\mathbf{t}}), \qquad \hat{\mathbf{Z}}_s \leftarrow \boldsymbol{\gamma}_b \odot \hat{\mathbf{Z}}_s + \boldsymbol{\beta}_b, \quad \hat{\mathbf{Z}}_f \leftarrow \boldsymbol{\gamma}_b \odot \hat{\mathbf{Z}}_f + \boldsymbol{\beta}_b. \tag{12}$$

After reshaping, both fused maps are upsampled to 28×28 to initialise the two decoders:

$$\mathbf{D}_s^0 = \mathrm{Up}_2(\hat{\mathbf{S}}_1), \qquad \mathbf{D}_f^0 = \mathrm{Up}_2(\hat{\mathbf{F}}_3). \tag{13}$$

3.7 Dual Decoders with Language-Guided Feature Fusion (LFFI)

Two symmetric decoders produce features at 28×28, 56×56, and 112×112, each conditioned on text via FiLM [23] at every stage (LFFI):

$$[\boldsymbol{\gamma}, \boldsymbol{\beta}] = \mathrm{MLP}(\bar{\mathbf{t}}), \qquad \mathrm{FiLM}(\mathbf{X}, \bar{\mathbf{t}}) = \boldsymbol{\gamma} \odot \mathbf{X} + \boldsymbol{\beta}. \tag{14}$$

At each scale, a decoder stage merges the current feature map with a skip connection:

$$\mathbf{Y} = \mathrm{ConvBlock}\Big(\mathrm{concat}(\mathbf{X}, \mathbf{S})\Big), \qquad \mathbf{Y}' = \mathrm{FiLM}(\mathbf{Y}, \bar{\mathbf{t}}), \qquad \mathbf{X}_{\mathrm{next}} = \mathrm{Up}_2(\mathbf{Y}'). \tag{15}$$

Semantic decoder. Uses semantic FPN skips $\mathbf{S}_{2:4}$:

$$\mathbf{D}_s^1 = \mathrm{Dec}_{28}(\mathbf{D}_s^0, \mathbf{S}_2), \quad \mathbf{D}_s^2 = \mathrm{Dec}_{56}(\mathbf{D}_s^1, \mathbf{S}_3), \quad \mathbf{D}_s^3 = \mathrm{Dec}_{112}(\mathbf{D}_s^2, \mathbf{S}_4). \tag{16}$$

Frequency decoder. Uses frequency encoder skips $\{\mathbf{F}_{\text{stem}}, \mathbf{F}_1, \mathbf{F}_2\}$:

$$\mathbf{D}_f^1 = \mathrm{Dec}_{28}(\mathbf{D}_f^0, \mathbf{F}_2), \quad \mathbf{D}_f^2 = \mathrm{Dec}_{56}(\mathbf{D}_f^1, \mathbf{F}_1), \quad \mathbf{D}_f^3 = \mathrm{Dec}_{112}(\mathbf{D}_f^2, \mathbf{F}_{\text{stem}}). \tag{17}$$

3.8 Output Heads and Fusion

Each decoder produces a single-channel logit map via a 1 × 1 convolution: $\mathbf{L}_s = \mathrm{Head}_s(\mathbf{D}_s^3)$ and $\mathbf{L}_f = \mathrm{Head}_f(\mathbf{D}_f^3)$. A learnable fusion head combines them:

$$\mathbf{L} = \mathrm{Head}_{\text{fuse}}\big(\mathrm{concat}(\mathbf{L}_s, \mathbf{L}_f)\big), \qquad \mathbf{M} = \sigma(\mathbf{L}). \tag{18}$$

$\mathrm{Head}_{\text{fuse}}$ is a small two-layer conv block that learns complementary branch weighting.

3.9 Training Objective

The model is supervised with a combined Dice and binary cross-entropy loss:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{Dice}(\mathbf{M}, \mathbf{Y}) + \lambda\, \mathcal{L}_{BCE}(\mathbf{M}, \mathbf{Y}), \tag{19}$$

where $\mathbf{Y}$ is the ground-truth binary mask. Optional deep supervision on branch logits further stabilises training:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \alpha\, \mathcal{L}_{Dice}(\sigma(\mathbf{L}_s), \mathbf{Y}) + \beta\, \mathcal{L}_{Dice}(\sigma(\mathbf{L}_f), \mathbf{Y}), \tag{20}$$

with small α, β to avoid over-constraining intermediate heads.

Trainable components. The FPN adapter, TCA adapters, FFBI, both decoders, and fusion head are trained end-to-end. The BiomedCLIP ViT backbone is frozen by default; TCA adapter weights introduce fewer than 2% additional parameters relative to the total trainable parameter count.

4 Dataset and Implementation Setup

Datasets. We evaluate FreqMedCLIP on three medical imaging benchmarks spanning different anatomies and imaging modalities. **Brain Tumor Segmentation (BraTS 2020)** [20,1,19]: 369 subjects with high-grade glioma imaged with multimodal MRI (T1, T1c, T2, FLAIR). We extract 2D axial slices from 3D volumes and filter uninformative slices, yielding 8,000 training and 2,200 validation slices. Tumour sub-regions include necrotic core, peritumoral oedema, and enhancing tumour. **Breast Cancer Segmentation (CBIS-DDSM)** [10]: 1,566 full-field digital mammograms with binary lesion annotations. Region-of-interest crops (512 × 512) are resized to 224 × 224, yielding 3,400 training and

Table 1: Dataset statistics. All splits follow a 70/20/10 stratified partition.

Dataset	Modality	#Train	#Val	#Test	Task
BraTS 2020	MRI (multi-modal)	8,000	2,200	1,100	Tumour segmentation
CBIS-DDSM	Mammography (2D)	3,400	800	400	Lesion segmentation
LUNA16	CT (axial slices)	4,200	1,000	500	Nodule segmentation

800 validation examples. **Lung Nodule Segmentation (LUNA16)** [26]: CT volumes from the LUNA16 challenge with 3D nodule masks annotated by four radiologists. We extract 2D axial slices to obtain 4,200 training and 1,000 validation examples from 1,018 volumes. Table 1 summarises the dataset statistics.

Class imbalance is addressed through Dice loss and balanced sampling during training. All images are resized to 224×224 and normalised per modality (z-score for MRI, min-max for CT and mammography).

Implementation Details. Table 2 lists the hyperparameters used for all experiments unless otherwise stated.

Table 2: Implementation hyperparameters.

Parameter	Value
Framework	PyTorch 2.0
Hardware	4× NVIDIA A100 (40 GB), DistributedDataParallel
Input resolution	$224 \times 224 \times 3$
Batch size	4 per GPU (16 total), gradient accumulation ×2
Optimiser	AdamW
Learning rate	1×10^{-4} for trainable modules (BiomedCLIP backbone frozen)
LR scheduler	CosineAnnealingLR ($T_{\max} = 100$)
Maximum epochs	100 (typical convergence: 50–75)
Early stopping	Patience 10 on validation Dice
Loss weights	$\lambda = 1.0$, $\alpha = \beta = 0.4$ (deep supervision)
Inference	∼80 ms per image (single A100), 3-checkpoint ensemble

Evaluation Metrics. We report three complementary metrics. The **Dice coefficient** $\text{Dice} = 2|P \cap G|/(|P| + |G|)$ serves as the primary overlap metric, following standard practice in medical segmentation [17]. **Intersection-over-Union** $(\text{IoU}) = |P \cap G|/|P \cup G|$ provides a stricter overlap measure that penalises false positives more heavily. The **Hausdorff Distance** (HD) measures worst-case boundary deviation in millimetres and serves as the primary boundary quality metric:

$$\text{HD}(P, G) = \max\Big(\max_{p \in \partial P} \min_{g \in \partial G} d(p, g),\ \max_{g \in \partial G} \min_{p \in \partial P} d(p, g) \Big). \tag{21}$$

5 Analysis and Results

5.1 Baseline Methods

We compare FreqMedCLIP against four baselines that span the design space from single-stream to foundation-model approaches.

BiomedCLIP (Semantic-Only). Frozen BiomedCLIP ViT with a standard U-Net decoder and no frequency stream. This baseline establishes the performance ceiling for pure semantic processing and isolates the value added by frequency information.

Frequency-Only. A single ConvNeXt-Tiny frequency encoder on DWT inputs with a U-Net decoder and no semantic stream. This tests whether frequency features alone can support medical segmentation.

UNet-CLIP. Text-conditioned U-Net using CLIP embeddings for decoder conditioning without explicit frequency decomposition. This represents the standard text-guided single-stream design.

SAM-Med2D [3]. The medical adaptation of SAM, representing the current strong foundation model baseline for 2D medical segmentation.

5.2 Quantitative Results

Table 3 presents results across the three datasets. FreqMedCLIP achieves the highest performance on all metrics and datasets.

Table 3: Segmentation performance (Dice ± std / IoU ± std) on three benchmarks. All models use the same 70/20/10 data split. Results are averages over three random seeds. Bold: best per column.

	Brain Tumor		Breast		Lung Nodule		Average	
Method	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
BiomedCLIP (Sem.)	$0.812_{\pm.021}$	$0.686_{\pm.019}$	$0.798_{\pm.018}$	$0.664_{\pm.016}$	$0.805_{\pm.025}$	$0.673_{\pm.022}$	$0.805_{\pm.021}$	$0.674_{\pm.019}$
Frequency-Only	$0.731_{\pm.031}$	$0.587_{\pm.028}$	$0.715_{\pm.027}$	$0.568_{\pm.024}$	$0.728_{\pm.029}$	$0.583_{\pm.026}$	$0.725_{\pm.029}$	$0.579_{\pm.026}$
UNet-CLIP	$0.834_{\pm.019}$	$0.714_{\pm.017}$	$0.821_{\pm.016}$	$0.696_{\pm.015}$	$0.827_{\pm.022}$	$0.705_{\pm.019}$	$0.827_{\pm.019}$	$0.705_{\pm.017}$
SAM-Med2D	$0.851_{\pm.017}$	$0.739_{\pm.016}$	$0.839_{\pm.015}$	$0.721_{\pm.014}$	$0.844_{\pm.020}$	$0.731_{\pm.018}$	$0.845_{\pm.017}$	$0.730_{\pm.016}$
FreqMedCLIP	$\mathbf{0.873}_{\pm\mathbf{.016}}$	$\mathbf{0.779}_{\pm\mathbf{.014}}$	$\mathbf{0.861}_{\pm\mathbf{.013}}$	$\mathbf{0.759}_{\pm\mathbf{.012}}$	$\mathbf{0.866}_{\pm\mathbf{.018}}$	$\mathbf{0.767}_{\pm\mathbf{.015}}$	$\mathbf{0.867}_{\pm\mathbf{.016}}$	$\mathbf{0.768}_{\pm\mathbf{.014}}$

The gains vary across datasets, and the pattern tells us something useful. FreqMedCLIP leads most strongly on brain tumour segmentation (Dice 0.873), where complex multi-region boundaries between necrotic core, oedema, and enhancing tumour demand precisely the complementary strengths of semantic context and frequency edge sensitivity. On breast mammography (Dice 0.861), dense tissue backgrounds create ambiguous boundaries that the semantic stream alone handles poorly (0.798). The frequency stream disambiguates fine-grained transitions that exhibit weak but consistent intensity edges in the Haar sub-bands, while language guidance ("breast cancer lesion") suppresses false activations on benign dense tissue. Lung nodule segmentation (Dice 0.866) shows the narrowest gap over SAM-Med2D (+2.2%). Nodules tend to have relatively clean circular boundaries where frequency-only methods already perform reasonably. Here the main benefit of our design is preventing false positives on vessels and bronchial walls via semantic gating.

The frequency-only baseline (0.725 Dice) confirms that frequency information alone is not enough: without semantic context, the encoder activates on all image edges indiscriminately. The semantic-only BiomedCLIP baseline (0.805 Dice) confirms the complementarity hypothesis. The 6.2% absolute gap between single-stream semantic and our full dual-stream model justifies the modest computational overhead (55% additional inference time, see Table 6).

5.3 Ablation Studies

Table 4 reports component-wise ablations on the average Dice metric. Each column shows the change relative to the full FreqMedCLIP model when the named component is removed.

Table 4: Ablation results. Each column shows ΔDice when the named component is removed from the full model. All variants trained with the same protocol, reported over three seeds.

	Full Model	w/o FreqEnc	w/o FFBI	w/o LFFI	w/o Dual Dec	w/o TCA
Avg Dice	0.867	0.825	0.837	0.846	0.852	0.861
Δ	—	−0.042	−0.030	−0.021	−0.015	−0.006

The frequency encoder is the single most important component (−4.2% without it), confirming that explicit boundary information from DWT sub-bands provides signal the semantic stream can't recover on its own due to attention smoothing. Removing FFBI costs 3.0%, showing that bidirectional cross-attention, which lets each stream selectively query the other, outperforms the naïve concatenation that removing FFBI reduces to. LFFI (−2.1%) validates that language-guided FiLM modulation at multiple decoder scales is effective beyond what the visual encoder alone provides: the text signal guides which frequency components get upweighted at each spatial scale, preventing semantic drift as resolution increases. The dual-decoder ensemble (−1.5%) and TCA adapters (−0.6%) provide smaller but consistent improvements.

5.4 Boundary Quality Analysis

Table 5 reports Hausdorff Distance and Boundary Dice, which are more sensitive to localisation precision than overlap metrics.

Table 5: Boundary quality metrics. HD (mm): lower is better. Boundary Dice: higher is better.

Method	Avg HD (mm)	Boundary Dice
BiomedCLIP (Sem.)	8.2 ± 1.1	0.742 ± 0.031
Frequency-Only	3.1 ± 0.5	0.701 ± 0.041
SAM-Med2D	6.1 ± 0.8	0.798 ± 0.025
FreqMedCLIP	$\mathbf{2.1 \pm 0.4}$	$\mathbf{0.831 \pm 0.020}$

FreqMedCLIP achieves 2.1 mm HD, a 3.9× improvement over the semantic-only BiomedCLIP baseline and 2.9× over SAM-Med2D. The frequency-only baseline achieves 3.1 mm HD (better boundary localisation than SAM-Med2D at 6.1 mm), yet its overall Dice is the lowest at 0.725. This dissociation is telling: precise boundary localisation without semantic context produces correct edges for the wrong objects. FreqMedCLIP resolves this by coupling both pathways. The FFBI module lets the semantic stream suppress spurious frequency activations on non-target structures, while the frequency stream sharpens boundaries that the semantic stream would otherwise blur.

5.5 Qualitative Results

Fig. 2 shows representative segmentation outputs for each dataset, and Fig. 3 visualises the per-stream activations after FFBI fusion.

Brain Tumor (BraTS) Breast (CBIS-DDSM) Lung Nodule (LUNA16)

Fig. 2: **Qualitative segmentation results.** One representative inference result per dataset. Ground truth boundary shown in green, FreqMedCLIP prediction in red, SAM-Med2D in blue. *[Placeholder: replace with actual inference outputs before submission.]*

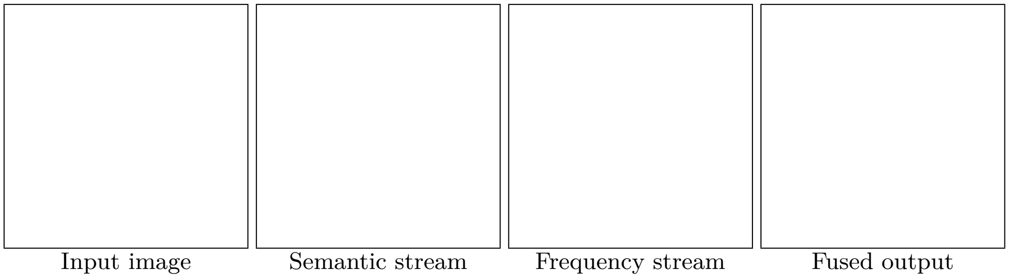

Fig. 3: **Per-stream activation maps.** Feature saliency after FFBI fusion. Semantic activations (centre-left) capture coarse anatomical regions. Frequency activations (centre-right) highlight boundary structures. The fused map (right) combines both, producing sharper, semantically constrained boundaries. *[Placeholder: replace with Grad-CAM or similar visualisations before submission.]*

5.6 Computational Efficiency

Table 6: Inference time and GPU memory on a single A100.

Method	Inference Time (ms)	Peak GPU Memory (GB)
BiomedCLIP (Sem.)	52	8.1
SAM-Med2D	156	12.3
FreqMedCLIP	81	10.7

FreqMedCLIP runs at 81 ms per image, faster than SAM-Med2D by 1.9× while achieving higher accuracy. The dual-stream overhead (29 ms relative to BiomedCLIP alone) comes primarily from the ConvNeXt-Tiny encoder. FFBI and the decoders add negligible latency. The 10.7 GB peak memory requirement fits within commonly available clinical workstation GPUs (A40, RTX 4090), making deployment practical.

6 Discussion

Interpreting the performance gains. The 6.2% absolute Dice improvement over the semantic-only BiomedCLIP baseline is large given BiomedCLIP's strong pretraining on fifteen million image-text pairs. We attribute this to two complementary mechanisms. First, DWT sub-band inputs give the frequency encoder explicit access to high-frequency image components that the ViT's patch-level processing systematically suppresses through attention smoothing, a documented frequency bias [22]. Second, FFBI bidirectional cross-attention lets the semantic stream ground its boundary predictions in actual edge signals rather than learned prior distributions over anatomical shape, reducing the blurring artefacts common in transformer decoders. The improvement over SAM-Med2D (+2.2% Dice, 2.9× better HD) deserves examination because SAM-Med2D uses substantially more pretraining data. The gap is most pronounced in boundary metrics, confirming that FreqMedCLIP's advantage lies in localisation precision rather than semantic recognition.

Clinical relevance of boundary quality. The 2.1 mm mean Hausdorff distance represents substantially tighter boundary localisation than the semantic-only baseline (8.2 mm) and SAM-Med2D (6.1 mm). In downstream workflows such as treatment planning and longitudinal lesion tracking, boundary errors propagate to volume estimates and margin decisions. The large HD reduction is therefore practically important even when Dice gains appear moderate. A formal clinical-impact analysis remains future work and would require task-specific prospective evaluation.

Frequency-semantic dissociation. The frequency-only baseline achieves lower HD (3.1 mm) than SAM-Med2D (6.1 mm) despite having dramatically lower Dice (0.725 vs. 0.845). This dissociation matters: SAM-Med2D produces segmentations that are semantically reasonable with smooth boundaries, while the frequency-only model draws sharp, localised edges that are geometrically precise when correct but frequently misattributed to non-target structures. This finding motivates the FFBI design choice directly. The semantic stream must suppress spurious frequency activations, not just concatenate with them.

Text guidance and its limits. LFFI provides a consistent 2.1% Dice improvement across all three datasets. Analysis of the ablation variants reveals that text guidance works best at intermediate scales (28×28, 56×56), where language conditioning helps the decoder commit to specific anatomical regions before refining fine details at 112×112. At the coarsest scale, the semantic stream already provides strong localisation. At the finest scale, spatial frequency features dominate. These observations align with the multi-scale language integration findings of LViT [13], though our FiLM-based implementation imposes less computational overhead than cross-attention conditioning.

Limitations. Three limitations deserve attention. *Resolution constraints*: the 224×224 input resolution inherited from BiomedCLIP limits detection of nodules smaller than about 20 pixels. ViT patch tokens at 14×14 spatial resolution provide insufficient coverage for sub-centimetre structures. Hierarchical multi-scale inference or higher-resolution ViT backbones would address this. *Diffuse boundaries*: while Haar DWT provides richer multi-directional sub-bands than a single Laplacian, diffuse infiltrative lesions (e.g., low-grade glioma) lack sharp intensity transitions and remain challenging for frequency-based boundary detection. Combining DWT with learned filterbanks may improve performance in these cases. *3D extension*: the current design operates on 2D slices, discarding inter-slice context present in volumetric MRI and CT. Extending to 3D requires rethinking both the DWT decomposition (3D sub-bands) and the attention scalability of FFBI.

7 Conclusion

We have presented FreqMedCLIP, a dual-stream architecture for text-guided medical image segmentation that couples a frozen BiomedCLIP vision transformer with a learned frequency encoder processing Haar wavelet sub-bands. The two streams interact at a shared bottleneck via bidirectional cross-attention (FFBI), and language-guided FiLM conditioning (LFFI) modulates both decoder branches at every spatial resolution. Experiments on brain, breast, and lung benchmarks show consistent improvements over strong baselines: 0.867 average Dice (vs. 0.845 for SAM-Med2D), 2.1 mm Hausdorff distance (vs. 6.1 mm for SAM-Med2D), and 12 M trainable parameters at 81 ms inference. Ablation studies confirm that the frequency encoder provides the largest individual contribution (+4.2% Dice), followed by FFBI fusion (+3.0%) and LFFI text guidance (+2.1%), with all components jointly necessary for the full performance. Future work will explore adaptive frequency decomposition with learnable filterbanks, extension to 3D volumetric segmentation with axial DWT, and self-supervised pretraining that jointly learns semantic and frequency representations.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (2017)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)
3. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: SAM-Med2D. arXiv preprint arXiv:2308.16184 (2023), preprint; no peer-reviewed venue confirmed
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
5. Huang, Z., Zhou, Y.: Frequency-aware U-Net for imbalanced medical image segmentation. arXiv preprint arXiv:2505.17544 (2025), concurrent preprint
6. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
7. 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). pp. 4015–4026 (2023)

8. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: MedCLIP-SAM: Bridging text and image towards universal medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 643–653. Springer (2024)
9. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: MedCLIP-SAMv2: Towards universal text-driven medical image segmentation. Medical Image Analysis p. 103749 (2025)
10. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data **4**, 170177 (2017)
11. Li, C., Wang, B., Zhang, Z., Gao, Y., Shu, R., et al.: Unleashing the potential of SAM for medical adaptation via hierarchical decoding. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36 (2023)
12. Li, X., Wang, Y., Zhang, P.: DUWS-Net: Wavelet-based dual U-shaped spatial-frequency fusion transformer network for medical image segmentation. Pattern Recognition **152**, 110422 (2025)
13. Li, Z., Li, H., Li, Q., Wang, P.A.H., Zou, Q., Wang, D., Yu, L.: LViT: Language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 96–107 (2024)
14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
15. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986 (2022)
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
17. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**, 195–212 (2024)
18. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **11**(7), 674–693 (1989)
19. Mehta, R., Filos, A., Baid, U., et al.: QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation — analysis of ranking metrics and benchmarking results. Journal of Machine Learning for Biomedical Imaging **1**, 1–26 (2022)
20. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015)
21. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention U-Net: Learning where to look for the pancreas. In: Medical Imaging with Deep Learning (MIDL) (2018)
22. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (ICLR) (2022)
23. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
25. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241. Springer (2015)
26. Setio, A.A.A., Traverso, A., de Bel, T., Berens, M.S.N., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical Image Analysis **42**, 1–13 (2017)
27. 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). pp. 4085–4095 (2020)
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(2) (2025)