FreqMedCLIP-SAMv2 Detailed Pipeline

1. Inputs (Đầu vào)

Pipeline bắt đầu với hai loại dữ liệu đầu vào song song:

- **Medical Image (I):** Ảnh y tế 2D (X-ray, CT slice, MRI slice, Ultrasound).
  - *Shape:* `(H, W, 3)` (RGB) hoặc `(H, W, 1)` (Grayscale).
  - *Ví dụ:* Ảnh chụp X-quang phổi của bệnh nhân.
- **Text Prompt (T):** Văn bản mô tả bệnh lý hoặc cấu trúc giải phẫu cần phân đoạn.
  - *Format:* Chuỗi ký tự (String).
  - *Ví dụ:* "A chest X-ray showing consolidation in the right lower lobe indicative of pneumonia." (X-quang ngực cho thấy sự đông đặc ở thùy dưới phổi phải biểu hiện của viêm phổi).

2. Stage 1: Frequency-Aware Feature Extraction (Trích xuất đặc trưng nhận thức tần số)

Thay vì chỉ sử dụng một encoder thông thường, chúng ta sử dụng cơ chế "Smart Single-Stream" để lấy cả đặc trưng ngữ nghĩa (semantic) và đặc trưng chi tiết (detail).

2.1. Frequency Injection (Tiêm thông tin tần số)

- **Input:** Ảnh gốc I.
- **Operation:** Discrete Wavelet Transform (DWT).
  - Dùng bộ lọc Wavelet (ví dụ: Haar hoặc Daubechies `db1`) để phân tách ảnh.
  - Thu được 4 thành phần: I_{LL} (Low-freq), I_{LH}, I_{HL}, I_{HH} (High-freq/Details).
- **Process:**
  - Chỉ giữ lại các thành phần High-Frequency ($I_{HF} = [I_{LH}, I_{HL}, I_{HH}]$).
  - Chuẩn hóa kích thước I_{HF} để phù hợp với feature map của các lớp đầu Encoder.
  - Đưa qua một lớp `Conv2d` nhẹ (Adapter) để chuyển đổi chiều dữ liệu.

2.2. BiomedCLIP Encoder (Single Stream)

- **Backbone:** Vision Transformer (ViT) từ BiomedCLIP (đã fine-tune với DHN-NCE).
- **Luồng xử lý:**
  1. **Early Layers (1-4):** Ảnh đi qua các lớp đầu tiên.
     - Tại đây, trích xuất **HF Features (F_{HF})**.
     - *Fusion nhẹ:* Cộng (Add) hoặc Nối (Concat) thông tin I_{HF} từ bước 2.1 vào feature map này.

- *Output:* F_{HF} (Chứa thông tin biên dạng, kết cấu sắc nét).

2. **Deep Layers (5-12):** Feature tiếp tục đi qua các lớp sâu.
   - Tại lớp cuối cùng, trích xuất **LF Features (F_{LF})**.
   - *Output:* F_{LF} (Chứa thông tin ngữ nghĩa cao cấp: "đây là phổi", "đây là tim").

3. Stage 2: Coarse-to-Fine Fusion (Pha trộn từ Thô đến Tinh)

Giai đoạn này thay thế cơ chế M2IB đơn giản của MedCLIP-SAMv2 gốc bằng quy trình 2 bước để đảm bảo độ chính xác biên dạng.

3.1. Semantic Localization (Định vị Ngữ nghĩa - Coarse Map)

- **Input:**
  - Image Semantic Features (F_{LF}).
  - Text Embeddings (E_T) (từ Text Encoder của BiomedCLIP).
- **Operation: M2IB (Multimodal Information Bottleneck)**.
  - Tính toán Mutual Information giữa Text và Image Features.
  - Lọc bỏ các thông tin nhiễu không liên quan đến Text.
- **Output: Coarse Saliency Map (S_{coarse})**.
  - Bản đồ nhiệt độ phân giải thấp, chỉ ra vị trí đại khái của đối tượng.
  - *Đặc điểm:* Đúng vị trí nhưng biên bị mờ (blobby).

3.2. Frequency Refinement (Tinh chỉnh Tần số - Fine Map)

- **Input:**
  - Coarse Saliency Map (S_{coarse}) (dùng làm hướng dẫn - guidance).
  - HF Features (F_{HF}) (chứa chi tiết biên).
- **Operation: FFBI-inspired Refinement Module**.
  - **Gating:** Dùng S_{coarse} để "khoanh vùng" trên F_{HF} ($F_{HF\_masked} = F_{HF} \odot S_{coarse}$).
  - **Sharpening:** Kích hoạt các pixel có giá trị gradient cao (biên) trong vùng đã khoanh.
  - **Fusion:** Kết hợp lại để tạo bản đồ cuối cùng.
- **Output: Fine Saliency Map (S_{fine})**.
  - Bản đồ nhiệt sắc nét, bám sát biên vật lý của tổn thương.

4. Stage 3: SAM-based Segmentation (Phân đoạn với SAM)

Giai đoạn này tận dụng khả năng zero-shot segmentation mạnh mẽ của Segment Anything Model (SAM).

4.1. Prompt Generation (Tạo gợi ý thị giác)

- **Input:** Fine Saliency Map (S_{fine}).
- **Process:**
  1. **Thresholding:** Áp ngưỡng (Otsu) để tạo Binary Mask.
  2. **Post-processing:** Loại bỏ các đốm nhiễu nhỏ (Connected Component Analysis).
  3. **Bounding Box Extraction:** Vẽ hộp bao (BBox) quanh vùng tổn thương lớn nhất.
- **Output:** Visual Prompts (Box coordinates: $[x_{min}, y_{min}, x_{max}, y_{max}]$).

4.2. SAM Inference (Suy luận)

- **Input:**
  - Ảnh gốc I.
  - Visual Prompts (BBox).
- **Model:** SAM (Image Encoder + Mask Decoder).
  - Ảnh gốc đi qua SAM Image Encoder (chạy 1 lần).
  - Visual Prompt đi qua SAM Prompt Encoder.
  - Cả hai đi vào Mask Decoder.
- **Output: Zero-shot Segmentation Mask (M_{zero}).**

5. Stage 4: Weakly Supervised Refinement (Tùy chọn)

Nếu có tập dữ liệu training (nhưng không có nhãn pixel), ta dùng kết quả trên để tự huấn luyện lại model chuyên dụng.

- **Input:** Tập dữ liệu ảnh $\{I_i\}$ và các nhãn giả $\{M_{zero\_i}\}$ vừa tạo ra.
- **Process:**
  - Huấn luyện mô hình **nnU-Net** trên cặp dữ liệu (I_i, M_{zero_i}).
  - Sử dụng **Uncertainty Estimation** (Checkpoint Ensembling) để lọc bỏ các vùng nhãn giả không tin cậy trong quá trình training.
- **Final Output:** Mô hình Segmentation chuyên dụng (chạy nhanh, nhẹ hơn SAM, độ chính xác cao hơn do đã lọc nhiễu).

Tóm tắt Luồng Dữ liệu (Data Flow Summary)

1. **Input:** `Image + Text` -> **Smart Encoder** (BiomedCLIP + Wavelet).
2. **Features:**
   - `LF Features` (Ngữ nghĩa) <- Layer sâu.
   - `HF Features` (Chi tiết) <- Layer nông + Wavelet.
3. **Fusion:**

- `Text + LF ->` **M2IB** `-> Coarse Map` (Vị trí).
- `Coarse Map + HF ->` **Refinement** `-> Fine Map` (Biên nét).

4. **Prompting:** `Fine Map ->` **BBox**.
5. **Segmentation:** `Image + BBox ->` **SAM** `-> Final Mask .`