Summary of Experimental Results

(a) Segmentation Performance (Test Set, n=45 images, 181 regions)

Method	Prompt Type	Params	Dice	IoU	Std
SAM2 (Hiera-L)	Box + Neg Points	224M	0.555	0.408	0.193
SAM2 (Hiera-L)	Bounding Box	224M	0.553	0.407	0.195
MedSAM (ViT-B)	Box + TTA	93.7M	0.536	0.389	0.191
MedSAM (ViT-B)	Bounding Box	93.7M	0.522	0.375	0.189
SAM2 (Hiera-L)	Multi-Point (5)	224M	0.418	0.287	0.209
SAM2 (Hiera-L)	Centroid (1)	224M	0.338	0.236	0.263

(b) Finetuning Comparison

Strategy	Epochs	Trainable Params	Test Dice	vs Zero-Shot
Zero-Shot (baseline)	0	0	0.555	-
Focal Loss (full)	50	224M (100%)	0.372	-33%
BCE Loss (full)	100	224M (100%)	0.371	-33%
LoRA (r=8)	30	4.2M (2%)	0.355	-36%

(c) CLIP Classification Accuracy

Prompt Strategy	Source	Accuracy	Macro F1
LLM Text + Few-Shot	GPT-4 + Examples	44.4%	0.338
Manual Visual v2	Expert-written	42.2%	0.311
LLM Text + Optimized	GPT-4	35.6%	0.270
LLM VLM + Few-Shot	Gemini + Images	29.4%	0.220
Manual Jargon v1	Expert-written	23.3%	0.138
LLM VLM v1	Gemini + Images	8.3%	0.091