MedCLIPSeg: Probabilistic Vision-Language Adaptation for Data-Efficient and Generalizable Medical Image Segmentation

Supplementary Material

A. Datasets Overview

Our evaluation comprises a diverse collection of medical image segmentation datasets, covering six organs and five imaging modalities. We organize our benchmarks into three settings: *data efficiency*, *fully supervised*, and *domain generalization*. The **data efficiency** and **fully supervised** evaluation includes BUSI [1], BTMRI [14], ISIC [17, 66], Kvasir-SEG [36], QaTa-COV19 [18], and EUS [35], which collectively span ultrasound, MRI, dermatoscopy, endoscopy, and X-ray modalities. For **domain generalization**, we employ ten diverse datasets to provide out-of-domain (OOD) samples: BUSUC [33], BUSBRA [28], BUID [4], UDIAT [8], BRISC [24], UWaterlooSkinCancer [2, 27], CVC-ColonDB [65], CVC-ClinicDB [6], CVC-300 [68], and BKAI [55], each introducing distinct appearance shifts across imaging devices, acquisition protocols, and anatomical domains. This combination enables a systematic analysis of segmentation robustness across both intra- and cross-domain distributions. Dataset statistics, modalities, and split details are summarized in Table S1. Importantly, our method does not use the validation sets; however, other methods, such as LViT [50], rely on them during training to select the best checkpoints. In our framework, models are trained on the training split, and we select the last epoch checkpoint to evaluate on the test split. For **domain generalization**, the OOD datasets are *never seen during training*; we evaluate directly on their test sets without any finetuning or adaptation.

B. Computational Cost Analysis

Table S2 summarizes the computational complexity of all compared methods, including parameter footprint, FLOPs, and inference time. All measurements are computed on the same BUSI [1] test set under identical hardware conditions. Although our `MedCLIPSeg` framework typically employs a sampling strategy during inference, the computational cost reported in Table S2 corresponds to the configuration where we use *only a single sampled forward pass*. This ensures a fair, per-sample comparison with other methods. In general, `MedCLIPSeg` exhibits a fair, competitive computational profile with state-of-the-art segmentation performance.

C. Text Prompt Generation

We introduce a scalable strategy for **automated caption generation in unpaired datasets** without relying on vision–language models, detailed in Algorithm 1. Instead of requiring image–text pairs, we query a large language model *once per dataset* to produce a small set of generic caption templates with placeholders for attributes such as *class, location, number, shape*, and *color*. Using lightweight image and mask processing, these attributes are automatically extracted and filled into the templates, enabling "clinician-style" descriptive captions.

Algorithm 1 Text Prompt Generation

Inputs: `Images`, `Masks` (optional), `Labels` (optional)
Goal: Produce one caption per image using LLM templates + simple attributes
Step 1: Templates (once per dataset)
Query an LLM to write a few short templates with placeholders: {class}, {location}, {number}, {shape}, {color}.
Provide separate "normal" and "lesion" templates.
Step 2: Attribute extraction (per image)
for each image in `Images` **do**
  if a corresponding mask exists **then**
    Class: use label if available; otherwise "lesion" if uniform.
    Location: coarse region from mask (e.g., upper/lower/left/right/center).
    Number: count connected components (single/multiple).
    Shape: coarse shape cue (e.g., round/irregular).
    Color: overall brightness/tone relative to background.
  else
    Mark as "normal" (no lesion mask).
  end if
end for
Step 3: Fill templates (per image)
for each image **do**
  if normal **then**
    Choose a "normal" template; save caption.
  else
    Choose a "lesion" template; replace placeholders with extracted attributes; save caption.
  end if
end for
Output: A list of paired `(image, text, mask)` samples ready for training or evaluation..

Table S1. **Summary of medical datasets:** Overview of datasets used in the data-efficiency, fully supervised, and domain generalization benchmarks. For *data-efficiency* experiments, values in parentheses under *Train* and *Validation* indicate the number of samples corresponding to (10%, 25%, 50%) of the full splits; other sections report full counts.

Dataset	Train	Validation	Test	Modality	Organ
Data-Efficiency Evaluation					
BUSI [1]	(62, 156, 312)	(7, 19, 39)	78	Ultrasound	Breast
BTMRI [14]	(273, 684, 1,369)	(132, 330, 660)	1,005	MRI	Brain
ISIC [17, 66]	(80, 202, 404)	(9, 22, 45)	379	Dermatoscopy	Skin
Kvasir-SEG [36]	(80, 200, 400)	(10, 25, 50)	100	Endoscopy	Colon
QaTa-COV19 [18]	(571, 1,429, 2,858)	(142, 357, 714)	2,113	X-ray	Chest
EUS [35]	(2,631, 6,579, 13,159)	(175, 439, 879)	10,090	Ultrasound	Pancreas
Fully Supervised					
BUSI [1]	624	78	78	Ultrasound	Breast
BTMRI [14]	2,738	1,321	1,005	MRI	Brain
ISIC [17, 66]	809	90	379	Dermatoscopy	Skin
Kvasir-SEG [36]	800	100	100	Endoscopy	Colon
QaTa-COV19 [18]	5,716	1,429	2,113	X-ray	Chest
EUS [35]	26,318	1,758	10,090	Ultrasound	Pancreas
Domain Generalization					
BUSUC [33]	567	122	122	Ultrasound	Breast
BUSBRA [28]	1,311	282	282	Ultrasound	Breast
BUID [4]	162	35	35	Ultrasound	Breast
UDIAT [8]	113	25	25	Ultrasound	Breast
BRISC [24]	4,000	1,000	1,000	MRI	Brain
UWaterlooSkinCancer [2, 27]	132	0	41	Dermatoscopy	Skin
CVC-ColonDB [65]	20	0	360	Endoscopy	Colon
CVC-ClinicDB [6]	490	61	61	Endoscopy	Colon
CVC-300 [68]	6	0	60	Endoscopy	Colon
BKAI [55]	799	100	100	Endoscopy	Colon

D. Detailed Hyperparameters

All models were trained using UniMedCLIP ViT-B/16 [43] as the vision backbone and PubMedBERT [76] as the text encoder. We employed the Adam [44] optimizer with a learning rate of 3×10^{-4}, a batch size of 24, and a cosine annealing learning rate schedule. The segmentation objective combines Dice and binary cross-entropy losses with equal weighting ($\lambda_{Seg}\mathcal{L}_{Seg} = \lambda_{Seg}\mathcal{L}_{Dice} + \lambda_{Seg}\mathcal{L}_{BCE}$ with $\lambda_{\text{Seg}} = 0.5$), while the CLIP-based contrastive alignment term was weighted by $\lambda_{\text{SoftCon}} = 0.1$. The probabilistic attention weighting factor was fixed at $\beta = 2.35$ across all experiments. All runs were performed on a single NVIDIA A100 GPU (40 GB). Due to the relatively large size of the EUS dataset, we observed a convergence within the first 10 epochs. Consequently, EUS experiments were limited to 10 epochs, whereas all other datasets were trained for 100 epochs to ensure full convergence under both data-efficient and domain-generalization settings. No validation set was used, and the checkpoint at the last epoch was utilized.

E. Prompt Designs Overview

Table S3 provides an overview of the text prompt configurations used in our ablation experiments (see Section 4.5). Each design type represents a distinct linguistic variation that probes the model's sensitivity to descriptive accuracy, spatial specificity, verbosity, and potential contradictions. By comparing these designs, we evaluate how differences in text formulation, from concise and spatially informative prompts to noisy or underspecified ones, influence segmentation performance and generalization across datasets.

F. 3D applicability

`MedCLIPSeg` naturally extends to 3D segmentation without modifying the core method, when given a 3D VLM backbone. We showcase this by using M3D-CLIP [5] to replace the 2D image encoder with a 3D one. 3D segmentations are obtained by computing the dot product between the global text token and 3D voxel features, followed by trilinear interpolation for upscaling. We validate this on the *CHAOS CT Liver dataset* [39] following the M3D-Seg data split and achieve a DSC of **88.72%**, demonstrating that

Table S2. **Comparison of computational complexity between different methods.** Models using text or multimodal supervision are marked with a ✓ in the *"Text?"* column.

Model	Text?	Params. (M)	FLOPs (G)	Inf. Time (s)
UNet [63]	✗	14.8	50.3	0.55
UNet++ [85]	✗	74.5	94.6	0.81
DeepLabv3 [12]	✗	57.6	38.4	1.16
AttnUNet [56]	✗	34.9	101.9	0.77
nnUNet [34]	✗	19.1	412.7	1.55
Swin-UNet [9]	✗	82.3	67.3	1.38
TransUNet [11]	✗	105	56.7	1.22
LViT [50]	✓	29.7	54.1	1.74
Ariadne's Thread [81]	✓	44.0	49.8	2.39
CLIPSeg [51]	✓	1.1	66.8	1.35
DenseCLIP [61]	✓	89.7	66.7	1.50
ZegCLIP [86]	✓	10.6	67.6	1.68
SAN [72]	✓	8.2	90.0	1.46
MaPLe [42]	✓	7.1	66.9	1.45
MaPLe [42] + Decoder	✓	8.2	67.3	1.75
VLSM-Adapter [19]	✓	5.0	68.4	1.32
CausalCLIPSeg [13]	✓	57.2	158.3	4.39
CAT-Seg [16]	✓	34.8	69.7	2.34
`MedCLIPSeg` (Ours)	✓	**18.7**	**73.6**	**1.51**

`MedCLIPSeg` generalizes beyond 2D. We report the *average runtime per volume* over 20 test volumes in the last column of Table S4, confirming practical feasibility for 3D settings, with further 3D analysis left for future work.

G. Inference-time MC sampling cost

Table S4 shows that using 5–10 MC samples only marginally affects DSC and uncertainty estimates. 2D runtimes are reported as the *average inference time per batch of 32 images*, measured over 1,000 test images, all on a single NVIDIA A100 GPU (40GB RAM). This demonstrates suitability for practical clinical settings with substantially reduced computational cost compared to 30 MC samples. In endoscopic video settings requiring ∼25–30 FPS, a 5 MC samples configuration is practical for real-time inference.

H. Effect of $\lambda_{SoftCon}$

Figure S1 illustrates the effect of the patch-level contrastive weight $\lambda_{SoftCon}$ on domain generalization. We find that $\lambda_{SoftCon} = 0.1$ provides the optimal balance between in-distribution (ID) and out-of-distribution (OOD) performance, yielding the highest harmonic mean (HM) score. When $\lambda_{SoftCon} = 0$, the contrastive loss is removed entirely, leading to a degradation in both ID and OOD performance due to the absence of semantic alignment across image–text patches. Increasing $\lambda_{SoftCon}$ beyond 0.1 results in marginal performance drops, suggesting that excessive contrastive weighting can overconstrain the feature space. These results highlight the importance of moderate patch-level contrastive regularization in maintaining both semantic consistency and domain robustness.

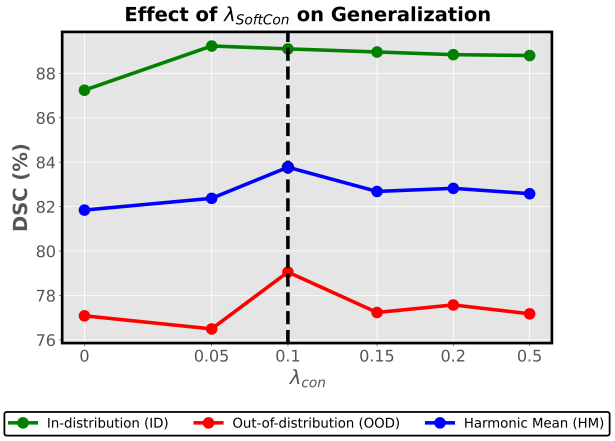

Figure S1. **Effect of $\lambda_{SoftCon}$ on Domain Generalization**

I. Effect of Gating Initialization

Table S5 examines the impact of the gating parameter initialization on segmentation performance across in-distribution (ID) and out-of-distribution (OOD) domains. We observe that initializing the gate with `sigmoid(0)` yields the best overall results, achieving the highest ID, OOD, and harmonic mean (HM) DSC scores. A smaller initialization value (`sigmoid(-0.5)`) makes the gate overly suppressive early in training, limiting information flow from the probabilistic branch and reducing generalization. Conversely, a larger initialization (`sigmoid(0.5)`) biases the fusion toward the probabilistic output too soon, leading to mild overfitting. The balanced initialization at `sigmoid(0)` thus provides a stable midpoint, enabling adaptive modulation between deterministic and probabilistic pathways throughout training.

J. Effect of "Two-way" Mechanism

Table S6 evaluates the contribution of the two-way cross-modal attention mechanism to segmentation performance. In the `Vision First` variant, cross-modal features are first computed for the vision tokens as the query in $\texttt{Attn}_{\texttt{PVL}}$ and then refined in the subsequent text-to-image interaction, while in `Text First`, this order is reversed. Among these, initializing fusion with the visual stream (`Vision First`) achieves the best results across in-distribution (ID), out-of-distribution (OOD), and harmonic mean (HM) DSC scores. Removing the two-way mechanism (`None`) or prioritizing text-driven conditioning (`Text First`) both lead to noticeable drops in OOD generalization, indicating that early visual grounding provides a stronger foundation for subsequent text-guided refinement. This suggests that

Table S3. **Prompt design taxonomy with examples.** Each configuration illustrates how wording choices (conciseness, spatial detail, contradictions, and noise) affect the semantics supplied to the model.

Design Type	Description Style	Example (Normal)	Example (Tumor)
`Original`	Balanced, accurate	"The breast appears normal with no signs of lesions."	"A malignant tumor is present in the upper-left region of the breast."
`Underdescriptive`	Minimal, label-only	"Normal breast."	"Tumor present."
`Overdescriptive`	Verbose, redundant	"The breast tissue appears entirely healthy, with homogeneous echotexture throughout."	"A clearly defined malignant tumor with irregular boundaries located in the upper-left quadrant."
`Contradictory`	Incorrect/Conflicting info	"Normal breast tissue with a visible lesion in the image."	"Malignant tumor detected, but breast appears completely normal."
`Missing Location`	No spatial info	"The breast appears normal with no signs of lesions."	"A malignant tumor is detected in the breast."

Table S4. **Performance-cost tradeoff under MC sampling**

Samples	Runtime (s/batch)	FPS	HM DSC (%)	HM Spearman (%)	3D Runtime (s/vol)
5	1.78	24.92	83.52	83.07	0.98
10	2.20	14.35	83.66	83.44	1.03
20	4.08	7.64	83.71	83.52	1.97
30	6.01	5.23	83.76	83.84	2.90

Table S5. **Effect of the gating initialization**

Gating Init. (g)	ID DSC (%)	OOD DSC (%)	HM DSC (%)
`sigmoid(-0.5)`	88.79	74.65	81.11
`sigmoid(0)`	**89.11**	**79.02**	**83.76**
`sigmoid(0.5)`	88.93	77.51	82.83

vision-first bidirectional fusion promotes more stable multimodal alignment, allowing the model to capture anatomical context before integrating semantic cues from text.

Table S6. **Effect of the two-way attention mechanism**

Two-way Mechanism	ID DSC (%)	OOD DSC (%)	HM DSC (%)
`None`	88.71	77.71	82.85
`Text First`	88.55	76.99	82.37
`Vision First`	**89.11**	**79.02**	**83.76**

K. Effect of Contrastive Pooling Mechanism

Table S7 analyzes the impact of different pooling strategies used in the contrastive loss. Among the three variants, `Average Pooling` achieves the highest in-distribution (ID), out-of-distribution (OOD), and harmonic mean (HM) DSC scores. This indicates that averaging patch-level embeddings provides a more balanced and stable global representation for contrastive learning compared to `[CLS]` or attention-based pooling. Removing uniform averaging (as in `Attention Pooling`) leads to noisier supervision due to bias toward high-attention regions, while relying solely on the `[CLS]` token underutilizes spatial information critical for dense prediction. Thus, average pooling yields the most consistent global-text alignment and best domain generalization.

Table S7. **Effect of the pooling strategies**

Contrastive Pooling	ID DSC (%)	OOD DSC (%)	HM DSC (%)
`[CLS]`	88.89	78.28	83.25
`Attention Pooling`	88.73	75.60	81.64
`Average Pooling`	**89.11**	**79.02**	**83.76**

L. Effect of Upscaling Blocks

Table S8 examines how varying the number of upscaling layers in the decoder affects segmentation performance. Using two upscaling blocks yields the best balance across in-distribution (ID), out-of-distribution (OOD), and harmonic mean (HM) DSC scores. A single block (`1`) limits spatial resolution recovery, resulting in coarse boundary predictions, while deeper configurations (`3`) introduce oversmoothing and reduce OOD robustness. The two-block design thus offers the optimal trade-off between preserving fine structural details and maintaining stable feature generalization across domains.

Table S8. **Effect of the number of upscaling layers**

Num. Upscale	ID DSC (%)	OOD DSC (%)	HM DSC (%)
1	88.73	75.74	81.72
2	**89.11**	**79.02**	**83.76**
3	88.64	74.99	81.24

M. Effect of Adapter Dimension (D_s)

Table S9 evaluates the impact of the shared dimensionality in the probabilistic vision-language (PVL) adapters. The best performance is achieved at a dimension of 256, balancing both in-distribution (ID) and out-of-distribution (OOD) segmentation accuracy. Smaller adapter sizes (e.g., 64 or 128) underfit the cross-modal representations, limiting their ability to capture nuanced semantic alignments between visual and textual features. Conversely, excessively large dimensions (e.g., 512) tend to overfit the training distribution, slightly reducing OOD generalization. The 256-dimensional configuration thus provides the optimal trade-off between expressiveness and regularization.

Table S9. **Effect of the shared dimension in the *PVL* adapters**

Adapter Dim. (D_s)	ID DSC (%)	OOD DSC (%)	HM DSC (%)
64	87.76	74.63	80.66
128	88.68	76.44	82.11
192	88.93	76.01	81.96
256	**89.11**	**79.02**	**83.76**
512	88.56	77.96	82.92

N. Effect of Different Confidence-Weighted Attention Mechanisms

Table S10 examines the impact of incorporating uncertainty into the attention computation in different manners. Our difference-based formulation yields the best performance across in-distribution (ID), out-of-distribution (OOD), and harmonic mean (HM) DSC scores. This approach adjusts attention weights by directly penalizing high-variance (low-confidence) regions, encouraging the model to focus on more reliable feature correspondences. For comparison, the *weight scaling* variant applies an uncertainty-dependent multiplicative attenuation to the attention matrix:

$$A^{\text{scaled}} = \texttt{softmax}(S_\mu) \oslash \big(1 + \beta S_\sigma\big),$$

where $\oslash$ denotes element-wise division. This strategy offers slightly lower gains compared to the proposed difference-based approach, while omitting uncertainty entirely reduces robustness under domain shifts. These results highlight that explicitly encoding confidence into attention promotes more stable and trustworthy segmentation performance.

O. Error vs. Uncertainty Correlation

The model's uncertainty maps exhibit a strong correlation with segmentation errors, as shown in Fig. S5. Across all ID and OOD datasets, the Pearson correlations between uncertainty and error are consistently high: 0.9248 (Breast Ultrasound), 0.9921 (Polyp Endoscopy), 0.9201 (Skin Dermatoscopy), and 0.9885 (Brain MRI), all with $p < 0.001$. These strong correlations indicate that regions of elevated uncertainty reliably align with areas of higher prediction error, confirming that the uncertainty estimates meaningfully reflect model confidence and can support downstream tasks such as boundary refinement, error correction, and active learning.

Table S10. **Effect of confidence-weighted attention mechanism**

Mechanism	ID DSC (%)	OOD DSC (%)	HM DSC (%)
`None`	88.97	77.49	82.83
`Scaling`	88.92	78.78	83.54
`Difference`	**89.11**	**79.02**	**83.76**

P. Deterministic vs. Probabilistic Confidence

As shown in Fig. S2, the probabilistic variant of `MedCLIPSeg` demonstrates superior handling of both overconfidence and underconfidence compared to its deterministic counterpart. Explicitly modeling predictive uncertainty suppresses spurious activations in non-lesion regions (reducing false positives) while recovering missed lesion boundaries (reducing false negatives). This leads to more balanced confidence calibration, smoother segmentation contours, and lower combined error rates across diverse ultrasound datasets.

Q. Additional Visualization Examples

We further evaluate `MedCLIPSeg` under cross-domain conditions using polyp endoscopy and breast ultrasound datasets. As shown in Figs. S3 and S4, the model maintains strong segmentation quality despite noticeable domain shifts in texture, lighting, and instrument artifacts. Uncertainty maps remain concentrated along polyp and tumor boundaries, reflecting well-calibrated confidence and robust generalization to unseen endoscopic and ultrasound environments.

R. Effect of Supervised Segmentation

Table S11 highlights the critical role of supervised segmentation annotations in medical image segmentation. When we remove the segmentation loss $\mathcal{L}_{\text{Seg}}$ and train the model solely with the soft patch-level contrastive objective $\mathcal{L}_{\text{SoftCon}}$, performance collapses to $\sim 20\%$ DSC in-distribution and below 13% out-of-distribution. Although purely contrastive or self-supervised objectives can be sufficient for natural-image segmentation, where object boundaries are often distinct and semantic categories are well separated and diverse, they are fundamentally insufficient

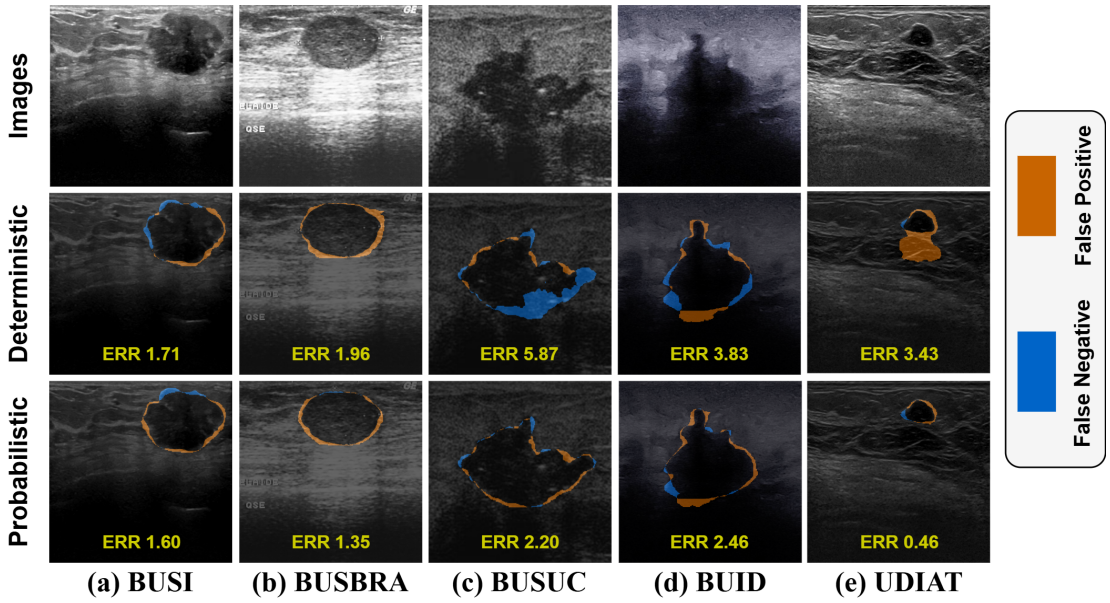

Figure S2. **FP/FN comparison between deterministic and probabilistic `MedCLIPSeg`.**

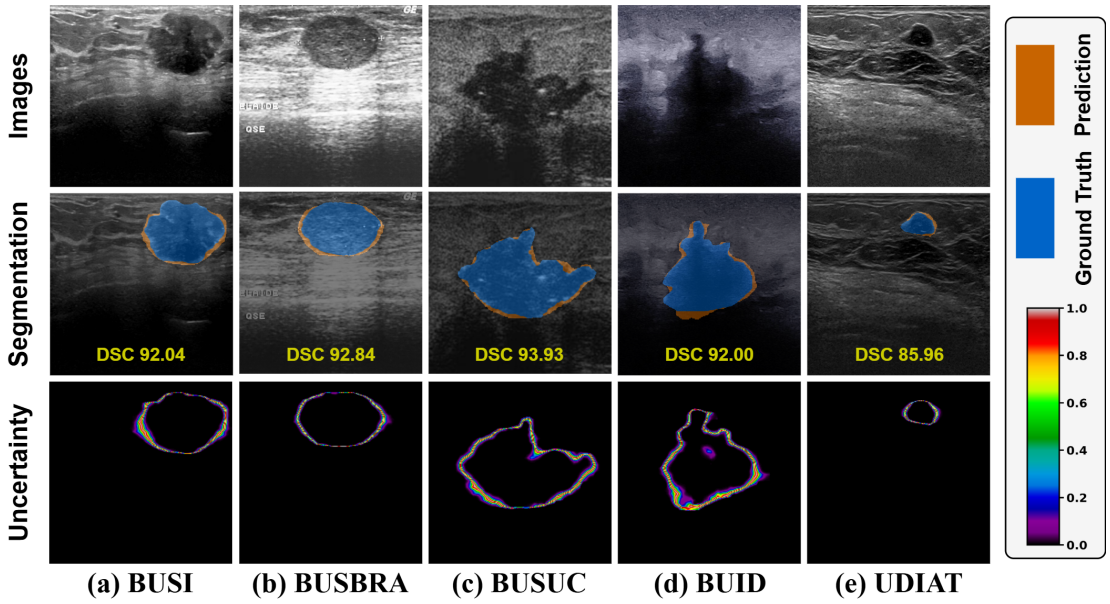

Figure S3. **Breast Tumor Ultrasound Segmentation and uncertainty visualizations.** Uncertainty peaks along lesion boundaries and remains consistent across breast ultrasound datasets, indicating reliable calibration and generalization.

in the medical domain. Medical boundaries are subtle, low-contrast, and frequently ambiguous, with fine-grained structures that require pixel-accurate supervision to disambiguate anatomy from imaging artifacts and surrounding tissues. Incorporating $\mathcal{L}_{\text{Seg}}$ provides this necessary spatial guidance, enabling the model to learn clinically meaningful decision boundaries and yielding dramatic gains of more than **+69% DSC** ID and **+66% DSC** OOD. These results clearly demonstrate that segmentation annotations remain indispensable for achieving reliable and generalizable medical image segmentation performance.

S. Additional Baselines

We include several additional baselines: a non-VLM nnU-Net with checkpoint ensembling [79], Ariadne's Thread [81], EviVLM [58], VLSM-Ensemble [20], and our framework's variant with deterministic adapters and an evidential

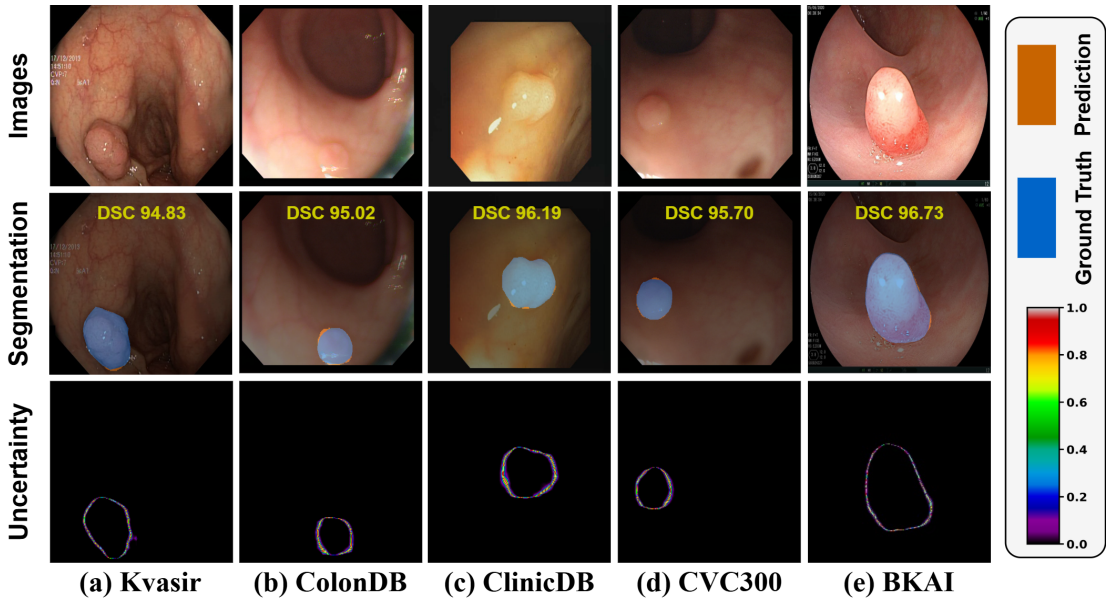

Figure S4. **Polyp Endoscopy Segmentation and uncertainty visualizations.** Uncertainty peaks along lesion boundaries and remains consistent across polyp endoscopy datasets, indicating reliable calibration and generalization.

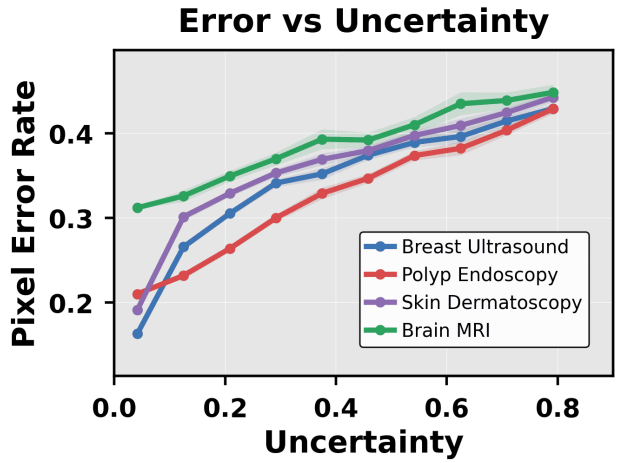

Figure S5. Relationship between predictive uncertainty and pixel-level segmentation error across imaging domains.

Table S11. **Effect of the segmentation annotations**

$\mathcal{L}_{Seg}$?	ID DSC (%)	OOD DSC (%)	HM DSC (%)
✗	19.84	12.68	15.47
✓	**89.11**	**79.02**	**83.76**

mask head at the end for pixel-wise uncertainty estimation without Monte Carlo (MC) sampling. We evaluate them all for robustness and uncertainty quality (**Spearman** correlation with errors) in Table S12, and show that `MedCLIPSeg` consistently outperforms the others in both aspects.

Table S12. **Domain generalization with more baselines (%)**

Method	ID		OOD		HM	
	DSC	Spearman	DSC	Spearman	DSC	Spearman
Ariadne's Thread [81]	68.25	–	27.24	–	38.94	–
EviVLM [58]	84.06	–	54.47	–	66.10	–
VLSM-Ensemble [20]	87.36	–	63.24	–	73.37	–
nnUNet + Ensembling [79]	86.50	66.74	74.20	55.22	79.80	60.44
MedCLIPSeg (Evidential)	88.18	78.49	76.61	76.79	81.99	77.63
`MedCLIPSeg` (Ours)	**89.11**	**87.57**	**79.02**	**80.41**	**83.76**	**83.84**

Table S13. **Domain generalization with different metrics (%)**

Method	ID	OOD	HM
SAN [68]	(86.9, 88.9, 84.5)	(74.3, 83.9, 69.9)	(80.1, 86.3, 76.5)
VLSM-Adapter [18]	(88.5, 87.9, 85.8)	(80.3, 80.8, 73.3)	(84.2, 84.2, 79.0)
CAT-Seg [15]	(87.2, 89.1, 86.1)	(81.1, 82.6, 74.6)	(84.0, 85.7, 79.9)
`MedCLIPSeg` (Ours)	(89.9, 90.8, 89.1)	(86.2, 80.7, 79.0)	(88.0, 85.5, 83.8)

T. Additional Evaluation Metrics

Table S13 reports **(Sensitivity, Specificity, F1)** for the top three baselines under domain generalization, supporting improved boundary localization by `MedCLIPSeg` under domain shifts.

U. Sample Text Prompts

Below, we provide one representative text prompt from each of the 16 datasets:

"one small pink round polyp, located in right of the image "

"A pituitary tumor is present in the center region of the brain."

"Presence of a benign lesion located at the upper section."

"Detected a malignant tumor positioned towards the center side."

"One small rectangle-shaped regular tumor at the left in the breast ultrasound image. "

"Findings indicate a benign tumor situated in the center area."

"no irregularities detected on MRI scan"

"one small white triangular polyp, located in center of the image"

"one small pink circle polyp, located in center of the image"

"one small white circle polyp, located in center of the image"

"Bilateral pulmonary infection involving two regions with involvement of all left lung and all right lung"

"Endoscopic ultrasound showing heterogeneous mass in center"

"one medium red rectangular skin melanoma which is a spot with dark speckles located in top right of the image "

"one medium white round polyp, located in left of the image"

"Detected malignant lesion located at the center area."

"Presence of a red skin melanoma positioned in the center part."

V. Per-dataset Efficiency Results

Tables S14, S15, S16, and S17 present detailed segmentation performance for each dataset across varying levels of labeled supervision (10%, 25%, 50%, and 100%). We report Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) scores to evaluate both volumetric overlap and boundary accuracy. This breakdown highlights the data-efficiency behavior of different model families, ranging from unimodal CNN and transformer baselines to text-driven and CLIP-based approaches. MedCLIPSeg consistently achieves the highest or second-highest performance across nearly all datasets and label fractions, demonstrating its robustness to annotation scarcity and strong cross-domain adaptability.

Table S14. **Per-dataset segmentation with 10% Labeled Data:** This table reports DSC and NSD values (%) across six medical image segmentation benchmarks. All baseline methods are trained using 10% of the ground-truth annotations. Best results are in **bold**, and second-best are underlined.

Method	BUSI		BTMRI		ISIC		Kvasir-SEG		QaTa-COV19		EUS	
	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
Unimodal Approaches												
UNet [63]	49.33	52.13	64.49	69.12	79.43	81.93	40.66	44.16	71.41	77.18	60.40	62.07
UNet++ [85]	53.80	57.58	62.23	66.11	82.83	85.54	46.27	49.28	69.22	74.99	67.96	69.00
DeepLabv3 [12]	42.45	45.64	61.96	66.47	84.62	87.72	45.14	48.32	68.51	74.38	65.22	66.50
AttnUNet [56]	54.66	58.17	58.68	62.77	85.16	88.08	42.46	45.68	70.86	76.37	64.85	66.44
nnUNet [34]	56.32	60.78	81.44	86.38	88.67	91.61	74.15	78.48	70.20	75.81	69.94	71.14
Swin-UNet [9]	39.87	45.20	41.26	45.83	81.04	84.12	36.84	43.05	62.10	70.43	57.14	58.81
TransUNet [11]	39.61	43.19	55.04	58.65	84.43	87.30	47.48	51.56	54.50	61.29	35.09	36.29
Generic Text-driven Approaches												
LViT [50]	63.37	65.97	52.48	54.80	74.53	75.48	51.60	53.10	76.52	82.23	80.57	81.21
Ariadne's Thread [81]	35.51	36.39	58.70	60.01	66.28	67.25	74.98	76.40	59.86	63.13	76.12	76.68
CLIP-Based Approaches												
CLIPSeg [51]	65.65	68.40	72.97	76.42	85.18	86.26	67.81	70.53	74.47	82.16	81.90	82.75
DenseCLIP [61]	55.09	57.41	60.56	61.87	88.54	89.56	73.73	75.80	66.95	73.86	62.18	63.48
ZegCLIP [86]	46.20	48.13	70.74	73.46	79.16	80.07	68.37	70.29	69.19	75.88	33.84	34.49
SAN [72]	66.99	69.56	77.92	81.76	89.20	90.21	66.79	69.36	72.36	78.78	71.51	72.17
MaPLe [42]	55.70	57.98	70.14	71.57	86.50	87.50	59.82	62.05	67.11	74.24	58.37	59.16
MaPLe [42] + Decoder	60.50	63.49	73.18	76.83	84.47	85.57	66.67	69.28	76.95	84.27	87.11	87.97
VLSM-Adapter [19]	63.85	66.60	73.19	76.81	86.81	87.95	71.74	74.44	74.90	82.06	76.31	77.14
CausalCLIPSeg [13]	51.29	53.02	73.97	77.27	84.89	85.86	60.35	62.24	70.58	77.18	82.61	84.17
CAT-Seg [16]	68.01	70.66	77.15	80.38	87.95	89.01	75.43	77.60	76.19	82.71	87.82	88.65
`MedCLIPSeg` (Ours)	68.66	71.35	79.07	82.71	90.35	91.40	77.21	79.53	79.73	86.27	91.59	92.38

Table S15. **Per-dataset segmentation with 25% Labeled Data:** This table reports DSC and NSD values (%) across six medical image segmentation benchmarks. All baseline methods are trained using 25% of the ground-truth annotations. Best results are in **bold**, and second-best are underlined.

Method	BUSI		BTMRI		ISIC		Kvasir-SEG		QaTa-COV19		EUS	
	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
Unimodal Approaches												
UNet [63]	56.28	59.90	70.40	74.98	82.44	85.68	41.88	44.93	68.23	73.05	57.23	58.45
UNet++ [85]	56.68	60.36	76.46	81.07	84.35	87.05	62.23	65.56	43.36	48.05	72.07	73.15
DeepLabv3 [12]	55.03	58.85	73.16	78.77	87.01	90.08	54.34	58.04	66.98	72.75	55.79	56.10
AttnUNet [56]	62.55	66.75	64.24	68.11	85.72	88.59	55.17	58.94	55.00	60.14	67.16	68.65
nnUNet [34]	62.28	66.29	83.20	89.12	89.85	92.76	80.96	84.91	72.84	78.41	71.26	72.44
Swin-UNet [9]	37.56	42.24	66.20	71.85	80.35	83.33	42.49	47.95	53.94	60.36	47.59	49.71
TransUNet [11]	46.21	49.91	58.33	62.24	86.04	88.82	51.51	55.95	50.09	56.19	39.33	40.57
Generic Text-driven Approaches												
LViT [50]	62.31	64.64	76.21	79.53	81.02	81.97	72.52	74.43	78.99	84.55	82.94	83.60
Ariadne's Thread [81]	39.01	40.04	58.70	60.01	68.53	69.45	75.71	76.11	60.06	64.29	76.54	77.14
CLIP-Based Approaches												
CLIPSeg [51]	70.35	73.18	74.91	78.40	86.45	87.59	73.67	76.41	78.77	85.74	85.72	86.72
DenseCLIP [61]	58.27	59.94	64.97	67.32	88.98	89.69	78.26	80.40	65.92	72.66	64.96	66.19
ZegCLIP [86]	49.86	51.88	73.18	76.11	79.39	80.24	71.95	73.75	72.99	80.09	87.37	88.00
SAN [72]	64.43	67.09	81.15	84.96	90.65	91.69	77.48	79.76	74.35	80.57	68.73	69.40
MaPLe [42]	63.94	66.42	72.87	73.91	87.89	88.89	71.68	73.85	67.43	74.68	65.39	65.93
MaPLe [42] + Decoder	67.12	69.75	79.78	83.74	87.89	88.98	74.96	77.57	79.52	86.15	88.58	89.39
VLSM-Adapter [19]	63.48	66.17	79.50	83.33	89.44	90.52	76.45	78.93	78.14	84.67	78.74	79.54
CausalCLIPSeg [13]	57.76	59.67	76.15	79.57	85.98	86.93	68.57	70.35	76.17	82.95	83.86	84.54
CAT-Seg [16]	73.08	75.76	80.07	83.59	89.61	90.67	80.11	82.50	79.12	85.45	84.73	85.54
`MedCLIPSeg` (Ours)	77.73	80.29	83.93	87.69	91.00	92.04	84.21	86.46	81.83	88.01	91.79	92.61

Table S16. **Per-dataset segmentation with 50% Labeled Data:** This table reports DSC and NSD values (%) across six medical image segmentation benchmarks. All baseline methods are trained using 50% of the ground-truth annotations. Best results are in **bold**, and second-best are underlined.

Method	BUSI		BTMRI		ISIC		Kvasir-SEG		QaTa-COV19		EUS	
	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
Unimodal Approaches												
UNet [63]	60.46	64.38	81.47	86.06	86.85	89.93	72.12	75.51	66.36	71.15	62.38	63.80
UNet++ [85]	63.63	67.16	80.21	84.61	88.29	91.23	74.53	77.72	63.80	67.71	68.42	69.43
DeepLabv3 [12]	55.83	60.21	77.82	83.56	86.14	89.06	67.72	71.90	64.86	70.29	59.12	60.39
AttnUNet [56]	59.81	63.46	72.96	77.24	87.28	90.48	74.79	78.66	64.71	69.95	68.52	69.99
nnUNet [34]	68.15	71.97	85.30	91.18	90.38	93.21	83.45	87.30	74.44	79.89	71.42	72.50
Swin-UNet [9]	41.86	48.11	57.76	64.96	85.52	89.14	50.04	55.25	53.67	61.58	46.51	48.44
TransUNet [11]	41.95	45.95	62.60	67.63	86.43	89.27	54.78	59.41	52.98	58.79	32.58	34.73
Generic Text-driven Approaches												
LViT [50]	62.74	65.29	78.89	82.17	89.18	90.20	78.63	80.54	80.21	85.51	83.63	84.36
Ariadne's Thread [81]	48.30	49.35	61.76	63.19	68.45	69.37	76.24	77.66	63.00	65.31	76.12	76.66
CLIP-Based Approaches												
CLIPSeg [51]	71.37	74.15	75.63	79.09	88.47	89.57	76.03	78.58	80.17	87.05	86.12	87.02
DenseCLIP [61]	64.62	65.78	68.83	70.21	89.17	90.18	80.16	82.29	63.93	70.74	65.81	67.52
ZegCLIP [86]	63.76	65.79	73.54	76.35	80.58	81.47	74.98	76.83	74.90	82.17	89.47	90.21
SAN [72]	71.53	74.16	82.08	85.87	91.06	92.09	80.03	82.25	75.74	81.92	72.36	72.84
MaPLe [42]	65.58	68.06	74.13	75.66	88.51	89.51	76.56	78.71	70.15	77.35	72.68	73.42
MaPLe [42] + Decoder	73.70	76.59	81.87	85.49	89.79	90.89	79.95	82.43	80.57	87.15	90.39	91.30
VLSM-Adapter [19]	69.61	72.51	82.47	86.46	91.35	92.42	82.98	85.59	79.33	85.50	79.22	80.13
CausalCLIPSeg [13]	68.48	70.82	75.69	79.08	88.11	89.08	73.26	75.20	76.76	83.07	85.00	85.95
CAT-Seg [16]	72.95	75.54	83.48	85.14	90.43	91.48	83.85	85.17	80.87	87.17	88.32	89.17
MedCLIPSeg (Ours)	81.48	84.06	85.93	89.75	91.97	93.03	88.18	90.36	82.97	89.13	92.57	93.36

Table S17. **Per-dataset segmentation with 100% Labeled Data:** This table reports DSC and NSD values (%) across six medical image segmentation benchmarks. All baseline methods are trained in a fully supervised manner using ground-truth annotations. Best results are in **bold**, and second-best are underlined.

Method	BUSI		BTMRI		ISIC		Kvasir-SEG		QaTa-COV19		EUS	
	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
Unimodal Approaches												
UNet [63]	70.04	73.88	86.06	90.71	89.10	92.06	80.26	83.53	76.97	82.32	68.53	69.92
UNet++ [85]	67.54	71.15	83.30	87.94	89.36	92.17	84.81	88.07	73.52	78.23	72.08	73.15
DeepLabv3 [12]	63.02	67.18	83.49	89.31	76.34	80.24	82.00	85.73	69.63	75.67	65.21	66.37
AttnUNet [56]	62.65	66.21	85.24	89.76	89.00	92.03	78.59	81.79	73.57	78.71	68.76	70.13
nnUNet [34]	76.85	80.70	86.91	92.00	90.52	93.37	85.44	89.29	75.43	81.00	73.27	74.09
Swin-UNet [9]	50.37	56.13	65.34	72.51	88.10	91.00	58.69	63.87	69.69	72.43	57.96	59.95
TransUNet [11]	57.98	62.50	70.90	76.46	88.10	91.00	58.69	63.87	68.74	71.99	58.88	61.10
Generic Text-driven Approaches												
LViT [50]	75.32	77.99	81.41	84.80	91.21	92.22	85.29	87.30	82.31	87.80	84.53	85.23
Ariadne's Thread [81]	57.26	58.22	69.96	71.40	68.37	69.30	77.42	78.79	70.70	73.94	76.71	77.31
CLIP-Based Approaches												
CLIPSeg [51]	80.95	83.87	85.33	89.45	90.55	91.62	81.98	84.61	81.76	87.30	88.66	89.58
DenseCLIP [61]	71.85	74.39	70.30	72.34	89.29	90.32	79.32	81.37	65.84	72.72	68.52	70.17
ZegCLIP [86]	72.08	74.45	76.65	79.77	81.45	82.33	78.46	80.43	75.42	82.59	89.83	90.54
SAN [72]	77.99	80.75	85.27	89.14	91.39	92.41	83.16	85.23	76.81	82.88	75.07	75.67
MaPLe [42]	66.37	68.92	75.40	76.83	88.31	89.30	76.12	78.27	70.40	77.52	70.98	71.75
MaPLe [42] + Decoder	80.49	83.38	85.08	89.20	90.10	91.21	83.46	85.96	81.86	88.16	88.65	89.55
VLSM-Adapter [19]	80.90	83.71	85.03	89.01	91.30	92.38	85.89	88.34	81.15	87.10	78.82	79.76
CausalCLIPSeg [13]	76.11	78.70	81.71	85.30	89.47	90.46	78.77	80.79	75.67	82.37	86.30	87.59
CAT-Seg [16]	81.83	84.52	84.86	86.52	91.27	92.34	86.43	88.83	82.82	88.60	88.18	89.07
MedCLIPSeg (Ours)	85.72	88.35	88.03	91.78	92.54	93.58	90.15	92.32	83.41	89.17	92.11	92.89