EXPLICIT ADAPTATION REGULATION FOR DATA-EFFICIENT MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-efficient medical image segmentation requires accurate in-domain predictions and reliable transfer to unseen datasets from limited source-domain annotations. Existing approaches largely regulate adaptation through latent features, leaving the connection between the current pixel–text assessment and subsequent updates to be learned indirectly through task supervision. With supervision concentrated on the final mask, this relationship can remain weakly constrained and become dependent on source-specific patterns, limiting performance under scarce supervision and domain shift. To bridge this gap, we introduce Explicit Adaptation Regulated Segmentation (**ExpARSeg**), which makes local pixel-text states explicit regulators of representation adaptation and spatial prediction. Specifically, ExpARSeg employs Evidential Pixel-Text (**EPT**) Adapters to condition bidirectional updates within frozen vision and text encoders on local support and commitment. It further integrates Selective Spatial Rendering (**SSR**) to direct text-conditioned visual corrections toward low-commitment locations identified by the current margin, with initial residual responses guiding subsequent point selection. Experiments across four imaging modalities and ten held-out target datasets show that ExpARSeg improves sample efficiency and domain generalization over state-of-the-art baselines using implicit adaptation regulation.

1 INTRODUCTION

Data-efficient medical image segmentation aims to achieve accurate in-domain predictions and reliable transfer to unseen target datasets using limited source-domain annotations. Pretrained medical vision-language models offer a promising foundation for this goal through knowledge learned from image-text pairs (Li et al., 2024; Liu et al., 2024; Koleilat et al., 2026). Prior approaches predominantly embed adaptation regulation in latent representation learning with its dependence on the current pixel-text assessment established indirectly through task supervision, which we characterize as *Implicit Adaptation Regulation*. Despite advances in medical image segmentation (Cho et al., 2024; Kerssies et al., 2025; Gui et al., 2026), these methods can suffer substantial degradation as annotation budgets shrink, while generalization to unseen target datasets remains inconsistent under limited source-domain supervision (Pahk et al., 2025). Such sensitivity to annotation scarcity and distribution shift reflects a persistent limitation in the data efficiency of existing methods.

We trace this limitation from a learning perspective. Finite mask supervision constrains segmentation outputs but leaves the underlying adaptation behavior undetermined. Source-domain fitting can weaken transferable semantic structure (Zhang et al., 2024; Yang et al., 2024a; Fahes et al., 2024; Pahk et al., 2025) and favor correlations that fail under distribution shift (Yang et al., 2024b; Tian et al., 2025; Zhou et al., 2025). Such a Latent-Induced Adaptation pathway lacks an explicit connection between the current prediction and subsequent spatial correction, which motivates our *Explicit Adaptation Regulation* to address this gap by directly conditioning refinement on the current pixel-text state. As illustrated in Fig. 1, explicit adaptation regulation makes pixel-text assessment a direct basis for both representation adaptation and spatial prediction, with local support and commitment governing feature updates and the evolving segmentation margin guiding selective correction. Prescribing this relationship in the architecture reduces reliance on regulation inferred from scarce annotations and keeps subsequent updates tied to the current assessment on unseen target domains, providing a task-related inductive bias for data-efficient segmentation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

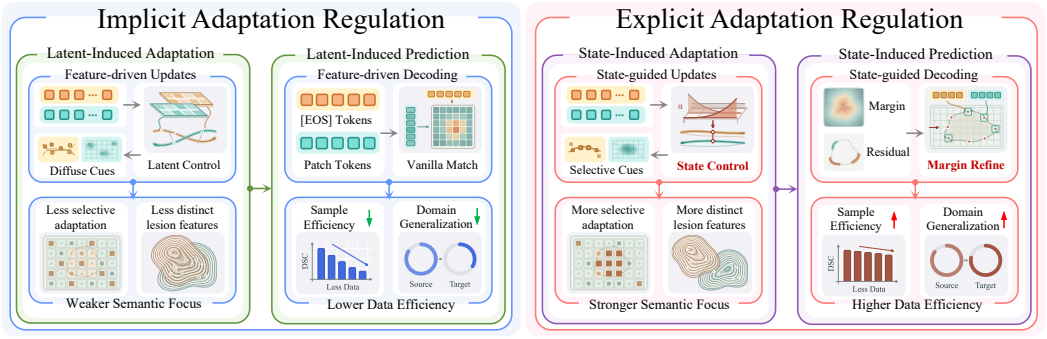

Figure 1: Comparison of implicit and explicit adaptation regulation, with the latter using pixel-text state to regulate adaptation and prediction for data-efficient medical image segmentation.

We instantiate this principle as **Expl**icit **A**daptation **R**egulated **Seg**mentation (**ExpARSeg**), a novel framework for data-efficient medical image segmentation. Specifically, ExpARSeg constructs evidential states from local pixel-text compatibility through **E**vidential **P**ixel-**T**ext (**EPT**) Adapters and uses these states to regulate bidirectional updates within frozen vision and text encoders. ExpARSeg further realizes state-induced prediction through a **S**elective **S**patial **R**endering (**SSR**) module, which applies the coarse pixel-text margin to select low-commitment locations for text-conditioned visual correction and incorporates half-resolution residual responses into full-resolution margin refinement. In summary, our main contributions include:

- We formulate explicit adaptation regulation as a design principle for data-efficient medical image segmentation, linking representation adaptation and spatial prediction to the current pixel-text assessment.
- We introduce EPT Adapters that derive evidential states from local pixel-text compatibility to better regulate bidirectional vision-language interactions.
- We design the SSR module to refine segmentation margins through text-conditioned visual residuals, with initial corrections guiding point selection across resolutions.
- We conduct experiments across four imaging modalities, multiple annotation budgets, and ten held-out target datasets. Results show that our ExpARSeg achieves superior performance than state-of-the-art baselines which rely on implicit adaptation regulation.

2 Related Work

Unimodal Medical Image Segmentation. Traditional medical image segmentation is predominantly built upon vision-only architectures, especially U-Net (Ronneberger et al., 2015) with the long skip connections to recover fine-grained spatial details. UNet++ (Zhou et al., 2018) further introduced nested and dense skip pathways to alleviate the semantic gap between encoder and decoder features. DeepLab series (Chen et al., 2017) employed atrous convolutions and atrous spatial pyramid pooling to capture multi-scale context. Attention U-Net (Oktay et al., 2018) incorporated attention gates to emphasize task-relevant regions. TransUNet (Chen et al., 2021) combines CNN-based local feature extraction with Transformer-based global context modeling. Swin-UNet (Cao et al., 2022) adopts hierarchical shifted-window self-attention to construct a fully Transformer-based U-shaped architecture. Beyond architectural design, nnU-Net (Isensee et al., 2021) adapted preprocessing, network configuration, training, and post-processing to individual datasets. Despite these advances, vision-only approaches are often sensitive to variations in image appearance, which limits their robustness to domain shifts caused by different scanners and imaging protocols.

Multimodal Medical Image Segmentation. Multimodal medical image segmentation exploits language to complement limited pixel supervision. Early medical-specific methods, such as LViT (Li et al., 2023) and Ariadne's Thread (Zhong et al., 2023), fused textual descriptions with visual features for lesion localization. In parallel, CLIP-based approaches adapted pretrained representations

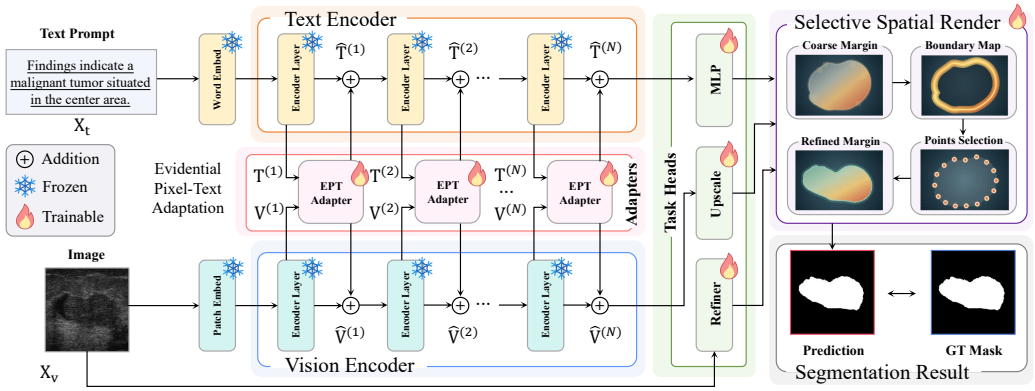

Figure 2: Overview of our proposed ExpARSeg. ExpARSeg connects frozen vision and text encoders through EPT Adapters, produces a coarse segmentation margin with task heads, and refines it through the SSR module to generate the final mask prediction.

to dense prediction through lightweight text-conditioned decoding (Lüddecke & Ecker, 2022), contextualized pixel–text matching (Rao et al., 2022), prompt adaptation (Khattak et al., 2023; Zhou et al., 2023), proposal generation with side networks (Xu et al., 2023), and encoder-only mask prediction (Kerssies et al., 2025). VLSM-Adapter (Dhakal et al., 2024) introduced lightweight adaptation blocks, whereas CausalCLIPSeg (Chen et al., 2024) employed cross-modal decoding and causal intervention to suppress confounding cues. CAT-Seg (Cho et al., 2024) formulated segmentation as the aggregation of dense image–text matching costs, providing a strong CLIP-based dense predictor. More recently, MedCLIPSeg (Koleilat et al., 2026) introduced probabilistic bidirectional vision–language interaction to model uncertainty during multimodal adaptation. Despite these advances, adaptation remains predominantly regulated implicitly through learned latent representations. Our method instead explicitly regulates both multimodal adaptation and spatial refinement using model-implied task states.

3 Method

3.1 Framework Overview

Fig. 2 illustrates the overall pipeline of our ExpARSeg. Given an image $\mathbf{X}_v$ and text prompt $\mathbf{X}_t$, ExpARSeg first extracts visual and textual token representations using the pretrained vision and text encoders. At layer n, an EPT Adapter constructs an evidential pixel-text state from the compatibility of visual tokens $\mathbf{V}^{(n)}$ and text tokens $\mathbf{T}^{(n)}$. The state regulates the following bidirectional interaction, producing adapted tokens $\widehat{\mathbf{V}}^{(n)}$ and $\widehat{\mathbf{T}}^{(n)}$ that are passed to subsequent encoder layers. Each adapter supplies residual updates to both frozen encoder streams. After the final layer N, lightweight task heads produce the coarse segmentation margin. SSR then refines this margin through text-conditioned visual corrections at low-commitment locations. During training, the pretrained vision and text encoders remain frozen, while the EPT Adapters, task heads, and SSR are jointly optimized. Implementation details appear in Appendix A.3–A.4.

3.2 Evidential Pixel-Text Adaptation

The EPT Adapter explicitly models pixel-text compatibility to regulate cross-modal adaptation. As shown in Fig. 3, it first updates visual tokens using text and then updates text using the adapted visual context. At an adapted layer n, let $\mathbf{V}^{(n)} \in \mathbb{R}^{(P+1)\times d_v}$ and $\mathbf{T}^{(n)} \in \mathbb{R}^{L\times d_t}$ denote P visual patch tokens with a global token and L text tokens, respectively. Learnable projections map both modalities to width d: $\mathbf{Z}_v = \mathbf{V}^{(n)}\mathbf{W}_v^{\downarrow}$ and $\mathbf{Z}_t = \mathbf{T}^{(n)}\mathbf{W}_t^{\downarrow}$. We omit the layer index below for clarity: index 0 denotes a global token and $p \in \{1, \ldots, P\}$ indexes visual patches.

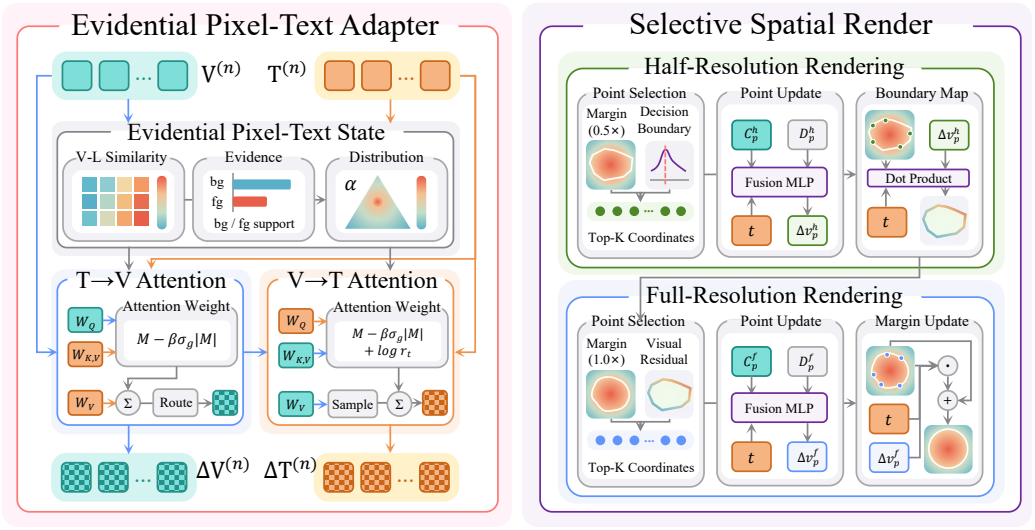

Figure 3: The architecture of (Left) the Evidential Pixel-Text (EPT) Adapter and (Right) the Selective Spatial Render (SSR). $W_{K,V}$ denotes the projection weights for attention keys and values.

Evidential Pixel-Text State. We construct a local compatibility score from each projected patch $\mathbf{z}_{v,p}$ and the global text token $\mathbf{z}_{t,0}$ as:

$$s_p = \tau \frac{\langle \mathbf{z}_{v,p}, \mathbf{z}_{t,0} \rangle}{\|\mathbf{z}_{v,p}\|_2 \|\mathbf{z}_{t,0}\|_2} + b, \qquad \tau > 0, \tag{1}$$

where τ and b are learned separately at each adapted layer. We represent the local pixel-text state with a binary Dirichlet distribution, a formulation also used for pixel-wise mask prediction (Schmidt et al., 2025). Specifically, we convert s_p into nonnegative background and foreground evidence:

$$\begin{aligned} e_{\mathrm{bg},p} &= \exp([-s_p]_+) - 1, \quad & e_{\mathrm{fg},p} &= \exp([s_p]_+) - 1, \\ \boldsymbol{\alpha}_p &= 1 + e_p, & S_p &= \alpha_{\mathrm{bg},p} + \alpha_{\mathrm{fg},p}, \end{aligned} \tag{2}$$

where $[x]_+ = \max(x, 0)$. The resulting distribution $\mathrm{Dir}(\boldsymbol{\alpha}_p)$ exposes the foreground probability ρ_p and uncommitted mass u_p:

$$\rho_p = \frac{\alpha_{\mathrm{fg},p}}{S_p} = \sigma(s_p), \qquad u_p = \frac{2}{S_p} = 2\sigma(-|s_p|), \tag{3}$$

where σ is the sigmoid function. This construction preserves $\log(\alpha_{\mathrm{fg},p}/\alpha_{\mathrm{bg},p}) = s_p$ and minimizes S_p among binary concentrations satisfying this log-ratio and $\alpha_{k,p} \geq 1$: one class retains unit concentration, while the other carries the evidence. Thus, u_p measures the lack of commitment implied by the current score. The state determines two directional routing factors:

$$r_{v,p} = 1 + (1 - u_p)(2\rho_p - 1), \qquad r_{t,p} = \max\{\epsilon,\ u_p + (1 - u_p)\rho_p\}\,. \tag{4}$$

The visual route strengthens foreground-supported updates and attenuates background-supported ones. The text route weights the visual context used to update language. Both routes approach one when s_p approaches zero. We also derive a shared stochastic gate from the same state:

$$g_p = 1 + \lambda\left(\theta_p - \frac{\alpha_{\mathrm{fg},p}}{S_p}\right), \qquad \sigma^2_{g,p} = \lambda^2 \frac{\alpha_{\mathrm{fg},p}\alpha_{\mathrm{bg},p}}{S_p^2(S_p + 1)}, \qquad \theta_p \sim \mathrm{Beta}(\alpha_{\mathrm{fg},p}, \alpha_{\mathrm{bg},p}), \tag{5}$$

where θ_p denotes a latent foreground probability associated with patch p, whose distribution is determined by the local pixel-text evidence. The scalar gate g_p modulates the features exchanged between modalities, while $\lambda \in (0, 1]$ controls the modulation strength and ensures that the gate remains positive. Subtracting the mean $\alpha_{\mathrm{fg},p}/S_p$ centers the gate at one, corresponding to identity scaling. The analytical variance $\sigma^2_{g,p}$ quantifies the gate variability, and its standard deviation $\sigma_{g,p}$

scales the attention-score penalty. Both directions share g_p and $\sigma_{g,p}$, deriving modulation from the same pixel-text state. Appendix B.1 provides a detailed analysis of this parameterization.

Text-to-Vision Attention. This branch introduces prompt semantics into visual representations. To account for the local pixel-text state, we discount cross-modal compatibility by the variability induced by its evidential gate. Using the projected visual and text tokens defined above, we compute

$$M_{pj}^{t\to v} = \frac{\langle \mathbf{z}_{v,p}\mathbf{W}_Q,\, \mathrm{LN}(\mathbf{z}_{t,j}\mathbf{W}_K)\rangle}{\sqrt{d}}, \qquad A_{pj}^{t\to v} = \mathrm{softmax}_j\big(M_{pj}^{t\to v} - \beta\sigma_{g,p}|M_{pj}^{t\to v}|\big)\,. \tag{6}$$

Here, $\mathbf{W}_Q$ and $\mathbf{W}_K$ are learnable query and key projections. LN denotes layer normalization. The term $\sigma_{g,p}|M_{pj}^{t\to v}|$ is the standard deviation of the gate-modulated score $g_p M_{pj}^{t\to v}$. The coefficient β controls the trade-off between compatibility and this variability penalty, favoring interactions whose compatibility remains strong after accounting for local modulation. The aggregated text features are then modulated by g_p and routed into the visual update:

$$\widetilde{\mathbf{z}}_{v,p} = r_{v,p}\left[(1-\eta_v)\mathbf{z}_{v,p} + \eta_v\mathcal{O}_v\left(g_p\sum_j A_{pj}^{t\to v}\,\mathrm{LN}(\mathbf{z}_{t,j}\mathbf{W}_V)\right)\right], \tag{7}$$

where $\mathbf{W}_V$ is the learnable value projection, $\mathcal{O}_v$ is the learnable output projection, and η_v balances the original visual representation and the attended text contribution. The final visual update is $\Delta\mathbf{V}^{(n)} = \widetilde{\mathbf{Z}}_v\mathbf{W}_v^{\uparrow}$, where $\mathbf{W}_v^{\uparrow}$ maps the update back to the vision encoder space.

Vision-to-Text Attention. This reverse interaction incorporates visual context into textual representations. We additionally apply $r_{t,p}$ as a spatial prior to account for the relevance of visual locations to the prompt. Text tokens now attend to the routed visual context $\widetilde{\mathbf{Z}}_v$:

$$\begin{aligned} M_{jp}^{v\to t} &= \frac{\langle \mathbf{z}_{t,j}\mathbf{W}_Q,\, \mathrm{LN}(\widetilde{\mathbf{z}}_{v,p}\mathbf{W}_K)\rangle}{\sqrt{d}}, \\ A_{jp}^{v\to t} &= \mathrm{softmax}_p\big(M_{jp}^{v\to t} - \beta\sigma_{g,p}|M_{jp}^{v\to t}| + \log r_{t,p}\big)\,. \end{aligned} \tag{8}$$

The first two terms apply the same penalty on score variability as above. Adding $\log r_{t,p}$ multiplies the unnormalized attention weights by $r_{t,p}$, downweighting locations with strong background support while retaining a near-neutral prior at low-commitment locations. The text update uses the shared gate to modulate visual contributions:

$$\widetilde{\mathbf{z}}_{t,j} = (1-\eta_t)\mathbf{z}_{t,j} + \eta_t\mathcal{O}_t\left(\sum_p A_{jp}^{v\to t} g_p\,\mathrm{LN}(\widetilde{\mathbf{z}}_{v,p}\mathbf{W}_V)\right). \tag{9}$$

Here, $\mathcal{O}_t$ and η_t are the corresponding text projection and mixing coefficient, respectively. The final textual update is $\Delta\mathbf{T}^{(n)} = \widetilde{\mathbf{Z}}_t\mathbf{W}_t^{\uparrow}$, where $\mathbf{W}_t^{\uparrow}$ maps the update back to the text encoder space.

3.3 Selective Spatial Rendering

The SSR module allocates spatial corrections according to the current segmentation margin. The final visual and global text features are normalized and passed through the upscaling head and text MLP, producing a quarter-resolution visual map $\mathbf{C} \in \mathbb{R}^{H/4\times W/4\times d_o}$ and a text vector $\mathbf{t} \in \mathbb{R}^{d_o}$. Their inner product defines the coarse margin $m_p^c = \langle \mathbf{C}_p, \mathbf{t}\rangle$. An extra lightweight image refiner extracts detail maps $\mathbf{D}^h$ and $\mathbf{D}^f$ from $\mathbf{X}_v$ at half and full resolution, respectively. Superscripts h and f denote these two rendering stages, and $\mathcal{I}_h$ and $\mathcal{I}_f$ denote bilinear resizing to their grids.

Half-Resolution Rendering. This rendering stage provides an initial correction at locations where the coarse prediction is weakly committed, using local image detail to complement the coarse visual context. To focus these corrections near the current decision boundary, we resize the margin to $m_p^{h,0} = \mathcal{I}_h(m_p^c)$ and select K_h locations with the lowest absolute margins:

$$\mathcal{P}_h = \mathrm{TopK}_p\big(-|m_p^{h,0}|; K_h\big)\,. \tag{10}$$

By Eq. (3), this selects the largest uncommitted masses $2\sigma(-|m_p^{h,0}|)$, concentrating updates near the current decision boundary (Kirillov et al., 2020). For each selected location, we obtain coarse

visual context $\mathbf{C}_p^h$ by bilinearly sampling the visual map $\mathbf{C}$ and extract the corresponding local detail feature $\mathbf{D}_p^h$ from $\mathbf{D}^h$. The fusion MLP combines these complementary features with a text-dependent interaction to predict a visual correction:

$$\Delta \mathbf{v}_p^h = \mathbf{W}_\uparrow^h \, \mathrm{MLP}_h\big([\phi_c^h(\mathbf{C}_p^h);\, \phi_d^h(\mathbf{D}_p^h);\, \phi_c^h(\mathbf{C}_p^h) \odot \phi_t^h(\mathbf{t})]\big)\,. \tag{11}$$

Here, ϕ_c^h, ϕ_d^h, and ϕ_t^h are learnable mappings that align the coarse visual feature, local detail feature, and text vector to a common feature space, respectively. $[\,;\,]$ denotes concatenation, and $\odot$ denotes element-wise multiplication. MLP_h denotes the half-resolution fusion MLP, and $\mathbf{W}_\uparrow^h$ projects its output back to the visual feature space. We collect local visual correction as:

$$B_p^h = \begin{cases} \langle \Delta \mathbf{v}_p^h, \mathbf{t} \rangle, & p \in \mathcal{P}_h, \\ 0, & p \notin \mathcal{P}_h. \end{cases} \tag{12}$$

Here, the text-projected boundary map expresses each visual residual in the same score space as the coarse prediction, allowing these localized correction cues to guide point selection at full resolution.

Full-Resolution Rendering. The half-resolution stage uses local image detail to update the spatial state for subsequent refinement. Projecting its visual residuals through the same text vector as the coarse prediction yields corrections in a common margin space. Appendix B.2 has analyzed spatial allocation, detail correspondence, and the half-resolution stage in detail. We upsample these corrections, add them to the coarse margin, and select K_f locations with low commitment:

$$\mathcal{P}_f = \mathrm{TopK}_p\big(-|\mathcal{I}_f(m_p^c) + \mathcal{I}_f(B_p^h)|; K_f\big)\,. \tag{13}$$

At each selected location, full-resolution detail $\mathbf{D}_p^f$ complements the visual context $\mathbf{C}_p^f$ sampled from the same coarse map $\mathbf{C}$. Their text-conditioned residual $\Delta \mathbf{v}_p^f$, computed using the full-resolution counterpart of Eq. (11), refines the margin as:

$$m_p^f = \begin{cases} \mathcal{I}_f(m_p^c) + \langle \Delta \mathbf{v}_p^f, \mathbf{t} \rangle, & p \in \mathcal{P}_f, \\ \mathcal{I}_f(m_p^c) + \mathcal{I}_f(B_p^h), & p \notin \mathcal{P}_f. \end{cases} \tag{14}$$

The final foreground probability is $\widehat{Y}_p = \sigma(m_p^f)$, where σ denotes the sigmoid function.

3.4 Training Objective

ExpARSeg combines segmentation supervision with an auxiliary target-conditioned KL (TC-KL) regularizer (Cai et al., 2025), whose contribution is analyzed in Appendix B.4. Through label conditioning, the regularizer penalizes evidence for incorrect classes in intermediate and final states:

$$\mathcal{L}_{\text{TC-KL}} = \langle D_{\mathrm{KL}}[\mathrm{Dir}(\widetilde{\boldsymbol{\alpha}}) \,\|\, \mathrm{Dir}(\mathbf{1})] \rangle\,, \tag{15}$$

where $\widetilde{\boldsymbol{\alpha}} = \mathbf{y} + (\mathbf{1} - \mathbf{y}) \odot \boldsymbol{\alpha}$ is the target-adjusted concentration, $\mathbf{y}$ is the corresponding one-hot label, and $\mathrm{Dir}(\mathbf{1})$ is the uniform Dirichlet reference. The brackets denote class-balanced averaging over the supervised states and spatial locations. Given a binary mask $\mathbf{Y}$ on the full-resolution grid Ω, the overall training objective is:

$$\mathcal{L} = \frac{1}{2}\mathcal{L}_{\mathrm{BCE}}(\widehat{\mathbf{Y}}, \mathbf{Y}) + \frac{1}{2}\mathcal{L}_{\mathrm{Dice}}(\widehat{\mathbf{Y}}, \mathbf{Y}) + \lambda_{\mathrm{KL}}\mathcal{L}_{\text{TC-KL}}, \tag{16}$$

where λ_{KL} is a hyperparameter that controls the strength of the auxiliary regularization.

4 Experiments

4.1 Experimental Setup

Datasets. We experiment on four medical segmentation tasks including Breast Ultrasound, Polyp Endoscopy, Brain MRI and Skin Dermatoscopy, each representing challenging and clinically critical applications with ambiguous boundaries. For supervised data-efficiency settings, we use BUSI (Al-Dhabyani et al., 2020), BTMRI (Cheng, 2017), Kvasir-SEG (Jha et al., 2019) and ISIC (Codella et al., 2018; Tschandl et al., 2018). For domain generalization settings, we use BUSUC (Iqbal & Sharif, 2024), BUSBRA (Gómez-Flores et al., 2024), BUID (Ardakani et al., 2023), UDIAT (Byra

Table 1: Quantitative comparison with state-of-the-art methods under different training-data ratios on four medical imaging modalities. Performance is reported in DSC (%). Best and second-best results are highlighted in **bold** and underline, respectively.

Method	Breast Ultrasound			Polyp Endoscopy			Brain MRI			Skin Dermatoscopy		
	10%	50%	100%	10%	50%	100%	10%	50%	100%	10%	50%	100%
UNet	62.34	68.01	77.23	53.43	78.15	80.37	72.53	83.95	86.36	85.98	88.99	89.39
UNet++	58.23	69.62	76.41	59.88	76.08	82.34	73.56	84.98	86.89	84.65	88.78	89.85
DeepLabv3	55.53	68.83	76.93	51.33	73.84	81.93	74.24	83.45	87.98	86.02	89.05	89.89
AttnUNet	53.95	67.77	76.10	49.69	76.53	83.31	73.51	84.24	86.03	84.26	89.52	89.76
nnUNet	48.65	68.76	79.60	66.62	80.48	84.60	75.21	83.60	88.32	86.06	89.67	90.37
Swin-UNet	49.13	61.15	67.95	47.59	58.16	73.17	60.65	76.56	81.86	84.75	88.35	89.64
TransUNet	47.43	64.72	76.40	51.47	67.82	77.56	67.12	80.98	84.83	84.70	89.12	90.30
LViT	62.23	72.83	80.27	43.12	71.68	80.88	70.25	83.33	87.24	81.68	87.12	89.60
Ariadne's Thread	54.37	75.71	79.91	52.96	75.28	80.94	75.15	83.06	85.63	80.45	89.19	89.83
EoMT-CLIP	69.21	79.10	83.11	76.58	82.77	81.81	79.95	83.31	84.88	87.54	89.58	90.70
CLIPSeg	65.48	76.86	79.87	66.04	76.63	79.93	76.16	81.03	83.44	85.80	89.52	90.47
DenseCLIP	57.53	69.29	79.63	76.43	80.73	87.09	78.65	81.38	83.79	90.56	91.86	91.53
ZegCLIP	63.22	70.08	77.35	53.88	75.07	76.29	73.91	79.32	80.60	82.31	87.04	88.71
SAN	43.29	29.26	51.31	57.56	58.43	64.23	33.57	53.99	55.96	81.57	76.38	79.87
MaPLe	52.76	68.41	72.36	70.34	80.61	83.17	64.58	79.86	81.31	87.54	90.07	90.26
MaPLe+ Decoder	54.71	66.19	75.42	70.84	81.36	82.77	70.54	81.94	83.86	87.78	90.03	91.03
VLSM-Adapter	65.79	74.65	78.39	64.78	73.01	75.75	74.09	78.40	80.71	84.84	88.20	89.79
CausalCLIPSeg	57.82	72.94	76.71	65.37	82.04	86.65	76.32	64.44	85.90	90.08	92.11	90.95
CAT-Seg	70.09	59.57	71.80	74.95	78.09	80.61	74.76	78.71	80.30	87.64	88.94	89.91
MedCLIPSeg	55.51	79.83	83.71	72.81	87.37	89.30	83.36	86.95	**88.74**	90.29	92.24	92.34
ExpARSeg (Ours)	**71.13**	**80.94**	**83.94**	**79.29**	**88.35**	**89.65**	**84.30**	**88.00**	88.67	**90.67**	**92.26**	**92.59**

Table 2: Quantitative comparison of domain generalization performance. Performance is reported in DSC (%). Src. and Tgt. denote performance on the source data and unseen target data from the same modality domain, respectively, while HM denotes their harmonic mean.

Method	Breast Ultrasound			Polyp Endoscopy			Brain MRI			Skin Dermatoscopy		
	Src.	Tgt.	HM	Src.	Tgt.	HM	Src.	Tgt.	HM	Src.	Tgt.	HM
UNet	77.23	59.89	67.46	80.37	52.38	63.42	86.36	71.40	78.17	89.39	47.48	62.02
UNet++	76.41	57.83	65.83	82.34	53.88	65.14	86.89	72.48	79.04	89.85	54.46	67.82
DeepLabv3	76.93	68.44	72.44	81.93	58.29	68.12	87.98	72.78	79.66	89.89	52.43	66.23
AttnUNet	76.10	57.55	65.53	83.31	59.03	69.10	86.03	71.76	78.25	89.76	51.30	65.29
nnUNet	79.60	70.90	75.00	84.60	67.69	75.21	88.32	71.41	78.97	90.37	65.52	75.97
Swin-UNet	67.95	61.59	64.61	73.17	41.20	52.72	81.86	66.68	73.49	89.64	48.12	62.62
TransUNet	76.40	40.94	53.31	77.56	57.12	65.79	84.83	71.24	77.45	90.30	45.65	60.65
LViT	80.27	33.89	47.66	80.88	51.79	63.15	87.24	65.31	74.70	89.60	40.15	55.46
Ariadne's Thread	79.91	42.57	55.55	80.94	49.37	61.33	85.63	69.64	76.81	89.83	33.71	49.02
EoMT-CLIP	83.11	62.69	71.47	81.81	65.65	72.85	84.88	77.15	80.83	90.70	38.14	53.70
CLIPSeg	79.87	46.84	59.05	79.93	61.94	69.79	83.44	73.36	78.08	90.47	26.99	41.58
DenseCLIP	79.63	73.49	76.44	87.09	72.91	79.37	83.79	66.21	73.97	91.53	67.56	77.74
ZegCLIP	77.35	18.05	29.27	76.29	52.43	62.14	80.60	70.40	75.15	88.71	27.60	42.10
SAN	51.31	34.89	41.53	64.23	40.36	49.57	55.96	37.53	44.93	79.87	42.85	55.78
MaPLe	72.36	48.13	57.81	83.17	63.70	72.14	81.31	64.32	71.82	90.26	41.38	56.74
MaPLe+ Decoder	75.42	56.39	64.53	82.77	65.88	73.37	83.86	70.51	76.60	91.03	64.49	75.50
VLSM-Adapter	78.39	48.35	59.81	75.75	56.14	64.49	80.71	69.69	74.80	89.79	25.00	39.11
CausalCLIPSeg	76.71	74.71	75.70	86.65	72.60	79.01	85.90	68.38	76.15	90.95	71.77	80.23
CAT-Seg	71.80	64.34	67.87	80.61	63.28	70.90	80.30	69.57	74.55	89.91	65.87	76.03
MedCLIPSeg	83.71	72.90	77.93	89.30	**78.99**	**83.83**	**88.74**	79.56	83.90	92.34	80.09	85.78
ExpARSeg (Ours)	**83.94**	**79.98**	**81.92**	**89.65**	78.59	83.75	88.67	**80.27**	**84.26**	**92.59**	**84.18**	**88.19**

et al., 2020), CVC-ColonDB (Tajbakhsh et al., 2015), CVC-ClinicDB (Bernal et al., 2015), CVC-300 (Vázquez et al., 2017), BKAI (Ngoc Lan et al., 2021), BRISC (Fateh et al., 2026) and UWaterlooSkinCancer (Amelard et al., 2014; Glaister et al., 2013). All datasets are publicly available with official data splits, segmentation masks, and text prompts, as detailed in Appendix A.1.

Implementation Details. All CLIP-based methods use UniMed-CLIP (Khattak et al., 2024) with a ViT-B/16 vision encoder and BiomedBERT (Gu et al., 2020) text encoder. EPT adapters follow blocks 1–10 in both frozen encoders, with $(\beta, \lambda) = (2.35, 1)$. SSR selects $K_h = 2048$ half-resolution and $K_f = 4096$ full-resolution points. We train the adapters, task heads, and SSR with Adam optimizer for 100 epochs at batch size 24, using cosine learning-rate decay from 3×10^{-4} to 10^{-4}). TC-KL uses $\lambda_{\mathrm{KL}} = 10^{-4}$ after a ten-epoch linear warm-up. Training shares one Beta sample per patch across both directions. Inference fixes $\mathbb{E}[g_p|\boldsymbol{\alpha}_p] = 1$, retaining the variance-based attention penalty. All experiments use four NVIDIA GeForce RTX 3090 GPUs (24 GB each). More information about detailed evaluation protocols are provided in Appendix A.5.

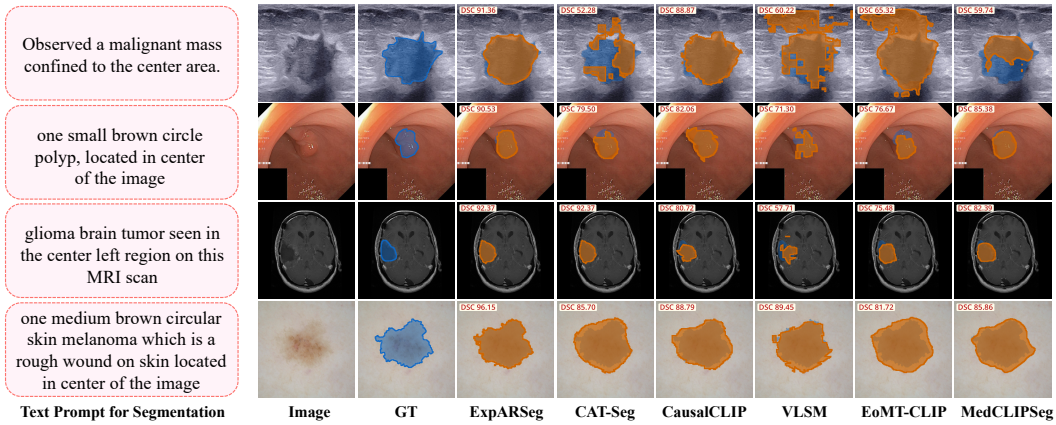

Figure 4: Segmentation visualization of our ExpARSeg and state-of-the-art baselines. Blue and orange represent mask annotation and model prediction, respectively.

Table 3: Component ablations in DSC (%). Source and Target average four and ten datasets, respectively. Subscripts give changes from Full (pp). Average summarizes all four scores.

Case	Method	Source (10%)	Source (50%)	Source (100%)	Target	Average
(a)	ExpARSeg (Full)	81.35	87.39	88.71	79.88	84.33
(b)	w/o Pixel-Text State	79.42 $_{(-1.92)}\downarrow$	86.08 $_{(-1.31)}\downarrow$	86.62 $_{(-2.10)}\downarrow$	70.32 $_{(-9.55)}\downarrow$	80.61 $_{(-3.72)}\downarrow$
(c)	w/o EPT Adapter	79.24 $_{(-2.11)}\downarrow$	86.39 $_{(-1.00)}\downarrow$	87.53 $_{(-1.19)}\downarrow$	75.35 $_{(-4.52)}\downarrow$	82.13 $_{(-2.20)}\downarrow$
(d)	w/o T→V Attention	68.27 $_{(-13.08)}\downarrow$	76.32 $_{(-11.07)}\downarrow$	77.94 $_{(-10.77)}\downarrow$	61.55 $_{(-18.32)}\downarrow$	71.02 $_{(-13.31)}\downarrow$
(e)	w/o V→T Attention	80.49 $_{(-0.85)}\downarrow$	87.18 $_{(-0.21)}\downarrow$	88.50 $_{(-0.22)}\downarrow$	49.41 $_{(-30.47)}\downarrow$	76.39 $_{(-7.94)}\downarrow$
(f)	w/o SSR	80.75 $_{(-0.60)}\downarrow$	86.78 $_{(-0.61)}\downarrow$	88.39 $_{(-0.32)}\downarrow$	77.84 $_{(-2.04)}\downarrow$	83.44 $_{(-0.89)}\downarrow$
(g)	w/o Half-Resolution Rendering	81.97 $_{(+0.62)}\uparrow$	86.66 $_{(-0.73)}\downarrow$	88.77 $_{(+0.05)}\uparrow$	77.72 $_{(-2.16)}\downarrow$	83.78 $_{(-0.55)}\downarrow$
(h)	w/o Full-Resolution Rendering	80.32 $_{(-1.03)}\downarrow$	86.95 $_{(-0.44)}\downarrow$	88.10 $_{(-0.62)}\downarrow$	78.25 $_{(-1.62)}\downarrow$	83.41 $_{(-0.93)}\downarrow$
(i)	w/o TC-KL Regularizer	81.65 $_{(+0.30)}\uparrow$	87.28 $_{(-0.11)}\downarrow$	89.29 $_{(+0.57)}\uparrow$	77.80 $_{(-2.07)}\downarrow$	84.01 $_{(-0.33)}\downarrow$

4.2 Performance Comparison

Sample Efficiency. We compare methods at different annotation budgets to assess learning from limited mask supervision. ExpARSeg leads 11 of 12 settings in Tab. 1. Its mean DSC across modalities exceeds the highest baseline mean by 3.03, 0.79, and 0.19 percentage points at 10%, 50%, and 100% annotation budgets, respectively. The larger gains at lower budgets support the benefit of explicitly regulating adaptation when supervision is scarce.

Domain Generalization. We evaluate transfer to unseen datasets within each modality to assess generalization beyond the source data. As shown in Tab. 2, ExpARSeg achieves the highest target DSC and HM in most cases. Averaged across modalities, it improves target DSC and HM over MedCLIPSeg by 2.87 and 1.67 points, respectively, compared with only 0.19 points in source DSC. These larger target gains with comparable source accuracy support improved cross-dataset transfer.

Segmentation Visualization. Fig. 4 compares lesion localization and boundary delineation across four modalities. ExpARSeg closely follows reference contours, while several baselines exhibit fragmented predictions or excessive foreground coverage, particularly in ultrasound and dermoscopy. These examples complement the quantitative gains by showing more complete lesion coverage and precise boundaries. More comparisons are provided in Appendix B.6.

4.3 Ablation Studies

Effect of Proposed Components. Tab. 3 evaluates the proposed components, where each variant is specified in Appendix A.6. Replacing state regulation throughout the pipeline reduces average DSC by 3.72 points in (b), while removing EPT alone reduces it by 2.20 points in (c). Within EPT, cases (d)–(e) show complementary roles, where text-to-vision attention mainly benefits source DSC and vision-to-text attention mainly improves transfer. Appendix B.3 examines this mechanism in detail. SSR improves target performance with two-stage rendering outperforming either stage alone.

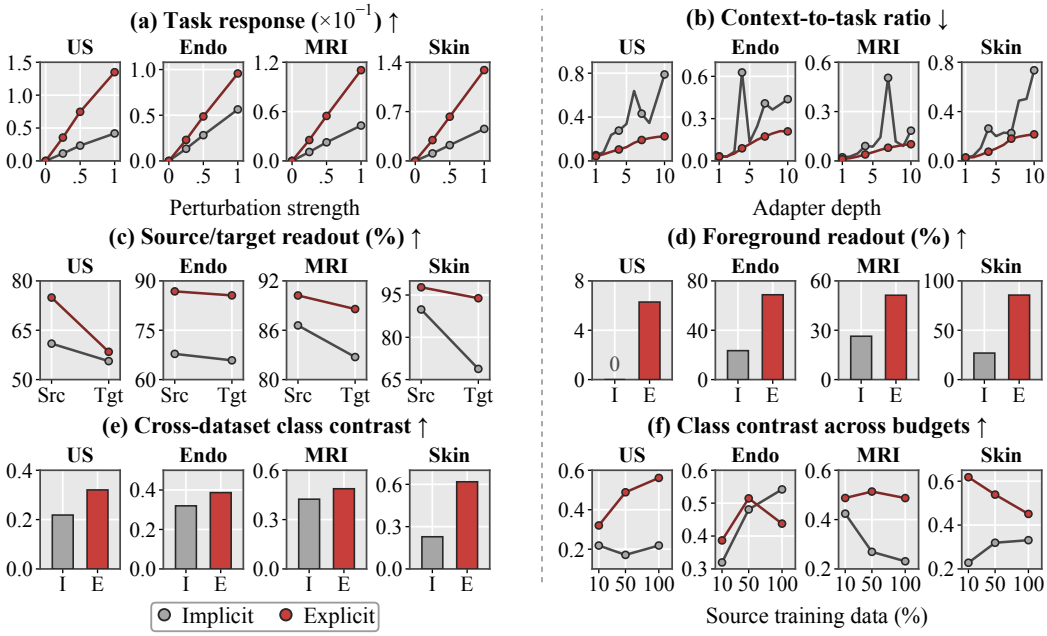

Figure 5: Analysis of our explicit mechanism. Task and context responses measure how adapter residuals near the lesion change when the lesion or background is blurred, respectively. Readout is the accuracy of a fixed source-trained classifier on adapted patch features. Class contrast denotes the cross-dataset cosine similarity gap within versus between lesion region and background.

Case (i) shows that TC-KL improves target DSC while source performance remains comparable, supporting its auxiliary role in transfer improvement.

Analysis of Explicit Adaptation Regulation. Fig. 5 examines our explicit mechanism with the interpretation detailed in Appendix A.7. Fig. 5(a) shows a larger change in ExpARSeg's visual adapter residuals near the lesion when lesion appearance is perturbed, measured by the ℓ_2 distance normalized by a common source feature scale. Fig. 5(b) compares responses to remote-background and lesion-region blurring at matched perturbation area and energy. Lower ratios at the final adapter indicate greater relative selectivity to lesion information. Fig. 5(c) shows higher class-balanced accuracy on source and target data when the same fixed source-trained classifier reads the features after the last shared adapter. Fig. 5(d) measures the fraction of target foreground patches correctly classified by this shared readout, with gains in every modality. Fig. 5(e) reports a larger gap between within-class and between-class cosine similarities across datasets, indicating clearer foreground-background separation in the feature space. Fig. 5(f) shows that this contrast advantage holds across all four modalities at 10% and 50% annotation budgets. Together, these results connect greater relative sensitivity to lesion changes with more accessible lesion information and clearer task structure across datasets. Layer-wise visualization in Appendix B.5 further illustrate how explicit regulation helps preserve lesion information in deeper representations.

5 Conclusion

In this paper, we propose ExpARSeg, a framework for data-efficient medical image segmentation based on explicit adaptation regulation. ExpARSeg consists of EPT Adapters to guide bidirectional feature updates with evidential pixel-text states and an SSR module to direct spatial refinement using prediction margins and visual residuals. Extensive experiments across four imaging modalities, multiple annotation budgets, and ten held-out target datasets validate the use of explicit states to guide adaptation under limited supervision. Moreover, it is worth noting that our state formulation does not explicitly distinguish visual ambiguity from errors in pixel-text alignment. Future research could explore more advanced adaptation paradigms that make intermediate updates intrinsically interpretable and establish conditions for guaranteed predictive improvement at each step.

AI USE STATEMENT

In this work, we used generative AI tools for refining hypotheses for mechanistic analyses, designing analysis experiments, implementing analysis procedures, translating text, and interpreting experimental results. We have not used generative AI tools for generating synthetic medical images or ground-truth segmentation masks, and theorem proving and proof writing are not applicable to this work. Additionally, we used generative AI tools for manuscript drafting and language polishing, literature searches, code debugging, figure preparation, and table and reference formatting. We have reviewed all AI-assisted work. Specifically, we checked the methodological descriptions against the implementation, tested the analysis scripts, and cross-checked plotted values and interpretations against the experimental records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper uses public medical datasets for methodological evaluation and does not involve any prospective patient recruitment or clinical intervention. Access to and redistribution of these datasets remain subject to the original providers' terms. Our proposed ExpARSeg is intended for research, and the reported benchmark results do not establish its suitability for clinical decision-making. Further clinical use would require independent validation, assessment of potential biases across patient populations and imaging conditions, and appropriate clinical oversight.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the method and experimental setup in the main text and provide additional implementation details in the appendix. The accompanying code is available in an anonymous repository at `https://anonymous.4open.science/r/ExpARSeg-E5EE/`. These materials complement the main paper with further context on the implementation and evaluation of ExpARSeg and are intended to facilitate independent reproduction and subsequent research.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Robert Amelard, Jeffrey Glaister, Alexander Wong, and David A Clausi. High-level intuitive features (hlifs) for intuitive skin lesion description. *IEEE Transactions on Biomedical Engineering*, 62(3):820–831, 2014.

Ali Abbasian Ardakani, Afshin Mohammadi, Mohammad Mirza-Aghazadeh-Attari, and U Rajendra Acharya. An open-access breast lesion ultrasound image database: applicable in artificial intelligence studies. *Computers in Biology and Medicine*, 152:106438, 2023.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Michal Byra, Piotr Jarosik, Aleksandra Szubert, Michael Galperin, Haydee Ojeda-Fournier, Linda Olson, Mary O'Boyle, Christopher Comstock, and Michael Andre. Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *Biomedical signal processing and control*, 61:102027, 2020.

Lile Cai, Chuan-Sheng Foo, Xun Xu, ZAIWANG GU, Jun Cheng, and Xulei Yang. Evidential learning-based certainty estimation for robust dense feature matching. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Yaxiong Chen, Minghong Wei, Zixuan Zheng, Jingliang Hu, Yilei Shi, Shengwu Xiong, Xiao Xiang Zhu, and Lichao Mou. Causalclipseg: Unlocking clip's potential in referring medical image segmentation with causal intervention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 77–87. Springer, 2024.

Jun Cheng. Brain tumor dataset. Dataset, 2017.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4123. IEEE, 2024.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Manish Dhakal, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Vlsm-adapter: Finetuning vision-language segmentation efficiently with lightweight blocks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 712–722. Springer, 2024.

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. A simple recipe for language-guided domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23428–23437, June 2024.

Amirreza Fateh, Yasin Rezvani, Sara Moayedi, Sadjad Rezvani, Fatemeh Fateh, Mansoor Fateh, and Vahid Abolghasemi. Brisc: Annotated dataset for brain tumor segmentation and classification. *Scientific Data*, 13(1):361, 2026.

Jeffrey Glaister, Robert Amelard, Alexander Wong, and David A Clausi. Msim: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis. *IEEE Transactions on Biomedical Engineering*, 60(7):1873–1883, 2013.

Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical physics*, 51(4):3110–3123, 2024.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

Jie Gui, Hang Tu, Wen Sha, and Xiuquan Du. Towards text-mask consistency in medical image segmentation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 55230–55247, 2026.

Ahmed Iqbal and Muhammad Sharif. Memory-efficient transformer network with feature fusion for breast tumor segmentation and classification task. *Engineering Applications of Artificial Intelligence*, 127:107292, 2024.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pp. 451–462. Springer, 2019.

Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan De Geus. Your vit is secretly an image segmentation model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25303–25313. IEEE, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122. IEEE, 2023.

Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9796–9805, 2020. doi: 10.1109/CVPR42600.2020.00982.

Taha Koleilat, Hojat Asgariandehkordi, Omid Nejatimanzari, Berardino Barile, Yiming Xiao, and Hassan Rivaz. Medclipseg: probabilistic vision-language adaptation for data-efficient and generalizable medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1406–1417, 2026.

Zhe Li, Laurence T. Yang, Bocheng Ren, Xin Nie, Zhangyang Gao, Cheng Tan, and Stan Z. Li. Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11704–11714, June 2024.

Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107, 2023.

Che Liu, Cheng Ouyang, Sibo Cheng, Anand Shah, Wenjia Bai, and Rossella Arcucci. G2d: From global to dense radiography representation learning via vision-language pre-training. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 14751–14773. Curran Associates, Inc., 2024. doi: 10.52202/079017-0471.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7076–7086. IEEE, 2022.

Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. In *International Symposium on Visual Computing*, pp. 15–28. Springer, 2021.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Jaehyun Pahk, Donghyeon Kwon, Seong Joon Oh, and Suha Kwak. Decoupled finetuning for domain generalizable semantic segmentation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 35935–35953, 2025.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18061–18070. IEEE, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Sebastian Schmidt, Julius Koerner, Dominik Fuchsgruber, Stefano Gasperini, Federico Tombari, and Stephan Günnemann. Prior2former - evidential modeling of mask transformers for assumption-free open-world panoptic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23646–23656, October 2025.

Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, and Jing Zhang. Black sheep in the herd: Playing with spuriously correlated attributes for vision-language recognition. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 33527–33560, 2025.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 180161, 2018.

David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15546–15561, 2023. doi: 10.1109/TPAMI.2023.3311618.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23826–23837, June 2024a.

Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael S. Yao, Chris Callison-Burch, James C. Gee, and Mark Yatskar. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 90683–90713. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2879.

Huatian Zhang, Lei Zhang, Yongdong Zhang, and Zhendong Mao. Homology consistency constrained efficient tuning for vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 93011–93032. Curran Associates, Inc., 2024. doi: 10.52202/079017-2953.

Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, and Ming Wu. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 724–733. Springer, 2023.

Nan Zhou, Huiqun Wang, Yaoyan Zheng, and Di Huang. Progressive parameter efficient transfer learning for semantic segmentation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 54352–54367, 2025.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11175–11185. IEEE, 2023.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pp. 3–11. Springer, 2018.

A Technical Details

In this section, we will provide a detailed description of important technical details that were not fully elaborated in the main text due to page limit.

A.1 Dataset Settings

We experiment on four medical modalities covering Breast Ultrasound, Polyp Endoscopy, Brain MRI and Skin Dermatoscopy. Each modality contains one source dataset for supervised training and several target datasets for out-of-distribution evaluation. All methods in the main comparison use the same training, validation, and test manifests at each annotation budget. It is worth noting that training and validation portions distributed with a target dataset are discarded.

Breast Ultrasound. BUSI (Al-Dhabyani et al., 2020) contains normal, benign, and malignant breast ultrasound images with binary lesion masks, retaining normal images with empty foreground annotations. BUSBRA (Gómez-Flores et al., 2024), BUSUC (Iqbal & Sharif, 2024), BUID (Ardakani et al., 2023), and UDIAT (Byra et al., 2020) provide four separate target collections to support a comprehensive assessment.

Polyp Endoscopy. Kvasir-SEG (Jha et al., 2019) provides annotated colonoscopy images for polyp segmentation. CVC-ColonDB (Tajbakhsh et al., 2015), CVC-ClinicDB (Bernal et al., 2015), CVC-300 (Vázquez et al., 2017), and BKAI (Ngoc Lan et al., 2021) form the four transfer targets. All five datasets contains binary masks, where polyp lesions are designated as the foreground.

Brain MRI. BTMRI (Cheng, 2017) denotes the prepared brain-tumor MRI collection used for supervised training, while BRISC (Fateh et al., 2026) is the held-out target including glioma, meningioma, pituitary tumors, and non-tumorous cases. We uses the released binary tumor masks.

Skin Dermatoscopy. ISIC (Codella et al., 2018; Tschandl et al., 2018) is the source for skin lesion segmentation, and UWaterlooSkinCancer (Amelard et al., 2014; Glaister et al., 2013) is the transfer target. Although these datasets share the segmentation task, their acquisition types should be distinguished: ISIC contains dermoscopic images, whereas the Waterloo collection contains dermatological photographs obtained with conventional cameras. This evaluation therefore includes a change in image acquisition, rather than a transfer restricted to two dermoscopic collections.

Text Prompts. We use the text prompts released with the benchmark, rather than generating new text for our experiments. The upstream prompt-construction procedure uses language-model templates populated with attributes extracted from images, masks, and available labels (Koleilat et al., 2026).

A.2 Input Processing

Input images are decoded as three-channel RGB arrays and resized directly to 224×224 using OpenCV bilinear interpolation. Grayscale images are represented by three channels. Mask annotations are loaded as grayscale images, resized with nearest-neighbor interpolation, and binarized at intensity 127. We set that intensities below 127 are background and those at or above 127 are foreground. The image tensor is scaled to $[0, 1]$ and normalized channel-wise with mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. All methods in our experiments use this same input transformation.

Data Augmentation. Our training pipeline applies a joint image-mask rotation with probability 0.5, drawing an integer angle uniformly from $[-20°, 20°]$. Rotation is performed on the resized images with the python PIL defaults, including nearest-neighbor resampling and zero fill.

A.3 EPT Adapter Implementation

For a batch of B images at 224×224, our EPT adapter receives visual tokens of shape $B \times 197 \times 768$, where 196 patches on a 14×14 grid and one global token. The text input has shape $B \times L \times 768$, where L is the tokenized sequence length with a maximum of 256. The first text token, `[CLS]`, supplies the global text feature.

Layer Configuration. Our EPT Adapter first uses separate bias-free linear projections to reduce the visual and text widths from 768 to 256. Each attention direction contains four biased $256 \to 256$ linear layers for query, key, value, and output projection, together with independent LayerNorm operations on keys and values. Attention is single-headed, and softmax is applied over the context-token dimension. Each direction also has a learned scalar mixing coefficient that combines its output-projected attention response with the incoming width-256 query features. Two separate bias-free $256 \to 768$ projections restore the encoder widths before residual addition.

A.4 SSR IMPLEMENTATION

SSR receives the normalized input image, a coarse visual feature map, and a text vector. The pretrained encoder projections produce 512-dimensional patch and global text features, which are separately ℓ_2-normalized before the task heads. The visual head reshapes the patches into $B\times512\times14\times14$ and applies two identical upscaling blocks, producing $\mathbf{C}\in\mathbb{R}^{B\times512\times56\times56}$. Each block contains a biased 2×2 transposed convolution with stride 2, GELU, a bias-free depthwise 3×3 convolution with padding 1, and channel LayerNorm. All layers preserve the 512-channel width. The text head contains three biased $512\to512$ linear layers with GELU after the first two, producing $\mathbf{t}\in\mathbb{R}^{B\times512}$. A batched inner product between $\mathbf{C}$ and $\mathbf{t}$ produces the coarse margin.

Image Refiner. The lightweight image refiner consists of a full-resolution path and a stride-2 half-resolution path. The full-resolution path applies a 3×3 convolution from 3 to 32 channels, followed by a depthwise 3×3 convolution with 32 groups. The half-resolution path applies a further 3×3 convolution from 32 to 32 channels with stride 2. All three convolutions use padding 1, have no bias, and are followed by eight-group GroupNorm and GELU. This yields detail maps of shape $B\times32\times224\times224$ and $B\times32\times112\times112$.

Point Rendering. The two stages use independently parameterized point networks that project coarse, detail, and text features to width 128 using linear layers followed by LayerNorm. Concatenating the projected coarse and detail features with their coarse–text interaction yields a 384-dimensional input to a $384\to128\to128$ MLP with GELU, followed by a zero-initialized $128\to512$ output projection. Its feature residual is dotted with $\mathbf{t}$ to obtain a margin correction. The half- and full-resolution stages select 2048 and 4096 points with the smallest detached absolute margins, respectively. Both sample the original coarse feature map at pixel centers and gather the corresponding detail features. Half-resolution corrections are upsampled and added to the coarse margin for full-resolution point selection. Selected full-resolution positions receive the coarse base plus the new correction, while other positions retain the interpolated half-resolution correction. Sampling uses border padding, and all spatial interpolation is bilinear with `align_corners=False`.

A.5 EVALUATION PROTOCOLS.

We mainly report DSC as the evaluation metric unless otherwise specified. For a predicted binary mask $\widehat{M}_i$ and annotation Y_i, the Dice similarity coefficient is

$$\mathrm{DSC}_i=\frac{2|\widehat{M}_i\cap Y_i|}{|\widehat{M}_i|+|Y_i|}. \tag{17}$$

Baseline Reproduction. We reproduce the baselines with reference to their official implementations, only making task-specific adaptations where needed for our medical image segmentation setting. All methods share the same data splits and evaluation protocol, and CLIP-based baselines share the same pretrained UniMed-CLIP backbone. Specifically, nnUNet (Isensee et al., 2021) uses a two-dimensional PlainConvUNet architecture under the shared split protocol, rather than the full automated planning, cross-validation, and ensembling pipeline. LViT (Li et al., 2023) and Ariadne's Thread (Zhong et al., 2023) consume the supplied descriptions through the integrated UTF-8 Transformer text encoder. MaPLe (Khattak et al., 2023) adapts both frozen towers with coupled prompts and uses dense pixel-text matching, while MaPLe+Decoder additionally includes an upscaling block. SAN (Xu et al., 2023) retains its bipartite set-prediction objective with auxiliary supervision and its polynomial learning-rate setup. Method-specific objectives and optimization settings are retained where required by official implementations.

Sample Efficiency. For each of the four source datasets, models are trained independently under the same 10%, 50%, and 100% manifest settings. Each checkpoint is evaluated on the same full source test set. Consequently, performance differences across budgets reflect the specified changes in source supervision rather than a change in the evaluation images.

Domain Generalization. We uses the 100% source checkpoint and applies it to every target paired within the same modality. Evaluation uses the target image and its supplied text prompt. Given "Src." is the source test score and "Tgt." is the arithmetic mean of the target-dataset scores within

that group. With source score S and target scores $D_1, \ldots, D_J$, we compute

$$T = \frac{1}{J}\sum_{j=1}^{J} D_j, \qquad \mathrm{HM}(S,T) = \frac{2ST}{S+T}. \tag{18}$$

We average dataset-level scores rather than pooling all target images, where each target dataset has equal weight. The MRI and skin groups each contain one target. The skin evaluation includes dermoscopy-to-photography transfer, as noted in Section A.1.

A.6 Ablation Protocols

Component Ablations. In Tab. 3, Case (b) removes pixel-text state in the whole pipeline, where EPT Adapters are replaced with ordinary bidirectional attention at the same positions and SSR is replaced with vanilla pixel-text similarity. Compared with Case (b), Case (c) retains the SSR module. Cases (d) and (e) retain only V→T attention and T→V attention within the EPT adapter, respectively. Case (f) only removes SSR and directly outputs the coarse margin. Case (g) uses only full-resolution rendering, selecting 4096 points from the resized coarse margin, while Case (h) uses only half-resolution rendering with 2048 points and then upsamples the complete rendered margin. Case (i) only removes the auxiliary TC-KL term, trained with segmentation loss alone.

A.7 Analysis Protocols

In the main text, Fig. 5 examines three complementary aspects of our proposed explicit regulation mechanism: whether local adaptation responds selectively to lesion-region changes, whether the resulting representations expose lesion information to a common readout, and whether foreground-background structure remains distinguishable across datasets. The Explicit denotes our full ExpARSeg, while Implicit denotes the variant in Tab. 3 Case (b). Panels (a)-(e) use 10% source annotations, while panel (f) also includes the 50% and 100% checkpoints.

Matched Regional Perturbations. Panels (a,b) compare perturbations near the lesion with perturbations in remote background on the target images. On the 224×224 annotation, square-neighborhood dilation by radii 2 and 8 defines the task region and a larger protected region, respectively. A deterministic geometry-based procedure selects a connected background region outside the protected region with the same area as the task region. We apply a 7×7 box blur within either selected region, leaving the prompt unchanged. If $\mathbf{d}_{i,A}$ is the RGB difference between the blurred and original image restricted to region $A \in \{\mathrm{task}, \mathrm{context}\}$, the intervention is

$$E_i = \min\left(\|\mathbf{d}_{i,\mathrm{task}}\|_2, \|\mathbf{d}_{i,\mathrm{context}}\|_2\right), \qquad \mathbf{X}_i^{A,\ell} = \mathbf{X}_i + \frac{\ell E_i}{\|\mathbf{d}_{i,A}\|_2}\mathbf{d}_{i,A}, \tag{19}$$

where $\ell \in \{0, 0.25, 0.5, 1\}$. Thus, the two conditions match both edited area and perturbation energy before the usual input normalization. Remote-background perturbations leave the entire protected region unchanged, and $\ell = 0$ provides an identity control. Images lacking a lesion, sufficient connected background, or nonzero common perturbation energy are excluded. Let $\Delta\mathbf{H}_{m,n,p}(\mathbf{X}_i)$ denote method m's visual adapter residual at depth n and patch p, and let $\mathcal{L}_i$ contain patches whose entire 16×16 support lies inside the protected region. We measure changes in these residuals using a common source-derived scale:

$$a_n = \mathrm{median}_{i,p}\, \|\mathbf{H}^{\mathrm{ref}}_{n,p}(\mathbf{X}_i) - \boldsymbol{\mu}_n\|_2, \tag{20}$$

$$R^A_{m,n}(i,\ell) = \frac{1}{|\mathcal{L}_i|}\sum_{p\in\mathcal{L}_i} \frac{\|\Delta\mathbf{H}_{m,n,p}(\mathbf{X}_i^{A,\ell}) - \Delta\mathbf{H}_{m,n,p}(\mathbf{X}_i)\|_2}{a_n}. \tag{21}$$

Here, $\mathbf{H}^{\mathrm{ref}}$ and its mean $\boldsymbol{\mu}_n$ come from the frozen, unadapted backbone on the 32 source fitting images; a_n is shared by both methods. *Task response* is R^{task} averaged over the ten adapter depths and then over images and target datasets. It quantifies how much the local visual update changes when lesion-region appearance is altered; the axis in panel (a) uses units of 10^{-1}. To distinguish stronger task response from generally larger responses, panel (b) measures

$$Q_{m,n}(i) = \frac{R^{\mathrm{context}}_{m,n}(i, 0.5)}{R^{\mathrm{task}}_{m,n}(i, 0.5)}. \tag{22}$$

We average these per-image ratios within each dataset and then across target datasets, separately at each depth. Neither responses nor depths are pooled before taking the ratio. A lower value indicates less influence from remote-background changes relative to matched lesion-region changes on the measured local update. Results show that Explicit has a larger task response in all four modalities and generally lower context-to-task ratios, including at the last adapter in every modality. Together, these observations support greater relative selectivity toward lesion-region information.

Shared Semantic Readouts. Panels (c,d) assess whether lesion information remains accessible through a common source-trained linear probe. For each source, we fit the probe on 32 images from the 10% training subset and select its regularization using 16 disjoint images from the same subset. Probe features come from the frozen, unadapted backbone at the depth corresponding to the last shared adapter. With source-fitting channel means $\boldsymbol{\mu}$ and standard deviations $\boldsymbol{s}$, we define

$$q(\mathbf{h}) = \sigma\left(\mathbf{w}^\top \frac{\mathbf{h} - \boldsymbol{\mu}}{\boldsymbol{s}} + b\right), \qquad \mathcal{L}_{\text{probe}} = \frac{1}{N_p} \sum_{i,p} \text{BCE}\left(q(\mathbf{h}_{ip}), y_{ip}\right) + \frac{\|\mathbf{w}\|_2^2}{2CN_p}, \tag{23}$$

where y_{ip} is foreground occupancy within patch p and N_p is the number of fitting patches. Channels with standard deviation at most 10^{-12} use scale one. The intercept is unpenalized, and optimization uses float64 L-BFGS for at most 150 iterations. The fixed probe reads each model's 768-dimensional features immediately after its last shared adapter, before the remaining encoder blocks. Features are collected deterministically, using mean gates for Explicit. Fixed evaluation cohorts contain at most 64 images per dataset. With $\mathcal{P}_{i,c}$ denoting the pure patches of class c and $\mathcal{C}_i$ the classes present in image i, we compute

$$A_{i,c} = \frac{1}{|\mathcal{P}_{i,c}|} \sum_{p \in \mathcal{P}_{i,c}} \mathbf{1}[\mathbf{1}[q(\mathbf{h}_{ip}) \geq 0.5] = c], \qquad A_i = \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} A_{i,c}. \tag{24}$$

Panel (c) reports image-averaged A_i on source and target data, with target-dataset means weighted equally within each modality. Panel (d) reports $A_{i,1}$ using the same aggregation on target images with pure foreground patches, directly measuring how often the probe recognizes lesion patches. The unchanged probe, normalization, and threshold ensure that differences arise from the representations presented to a common decision rule. Results show that Explicit improves both source and target readout in all four modalities. Averaged equally over the ten target datasets, overall readout increases from 63.74% to 75.86%, and foreground readout from 14.71% to 43.72%, while background readout exceeds 99% for both methods. Thus, the improvement includes greater accessibility of lesion information, beyond accurate recognition of abundant background patches.

Transferable Task Structure. Panels (e,f) examine whether the readout advantage is accompanied by clearer class structure across datasets, independently of a learned classifier. We extract 512-dimensional visual patch features after the final encoder block, normalization, and pretrained projection, before the task head and SSR. Each patch vector is ℓ_2-normalized. Using the same pure-class thresholds, we deterministically select at most 32 patches per image and class, retaining identical images and patch coordinates across methods and training budgets. Let $\bar{\mathbf{f}}_{i,c}$ denote their mean without further normalization, and let $\mathcal{I}_{a,c}$ contain images in dataset a with eligible class-c patches. The cross-dataset similarity is

$$S_{ab}^{cc'} = \frac{1}{|\mathcal{I}_{a,c}|\,|\mathcal{I}_{b,c'}|} \sum_{i \in \mathcal{I}_{a,c}} \sum_{j \in \mathcal{I}_{b,c'}} \bar{\mathbf{f}}_{i,c}^\top \bar{\mathbf{f}}_{j,c'}, \qquad a \neq b, \tag{25}$$

which averages patch-pair cosine similarities with equal weight per image pair. We define class contrast as the difference between same-class and different-class similarities:

$$C_{\text{class}} = \frac{1}{2|\mathcal{D}|} \sum_{(a,b) \in \mathcal{D}} \left(S_{ab}^{00} + S_{ab}^{11} - S_{ab}^{01} - S_{ab}^{10}\right), \tag{26}$$

where $\mathcal{D}$ includes all unordered cross-dataset pairs within a modality, including source-target and target-target pairs. Higher contrast indicates greater relative separation of foreground and background across datasets, which does not require higher absolute same-class similarity. Panel (e) shows higher contrast for Explicit in all four modalities at 10% supervision. Panel (f) repeats the measurement with models trained at 10%, 50%, and 100%, keeping the diagnostic samples fixed. Explicit achieves higher contrast on all 22 dataset pairs at both 10% and 50%, which supports the preservation of cross-dataset task structure under limited source supervision.

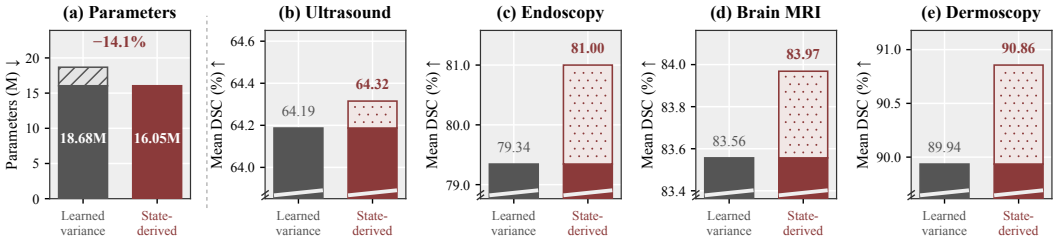

Figure 6: Comparison with latent variance modeling under 10% source annotations. Diagonally hatched regions indicate extra parameters required by learnable variance heads, while dotted regions show mean DSC gains from our state-derived explicit regulation.

Joint Analysis. In summary, the experimental results above can jointly support the proposed role of explicit regulation in directing adaptation toward lesion evidence while retaining task structure across datasets under limited source supervision. Specifically, greater relative sensitivity to lesion-region changes is accompanied by better lesion recognition under a fixed source-trained readout, connecting selective local responses to useful task information. The higher class contrast further shows that this improvement accompanies clearer foreground-background separation across datasets, extending the evidence beyond the probe's particular decision boundary.

B Supplementary Experiments

In this section, we will provide more supplementary experiments to further evaluate our ExpARSeg.

B.1 Comparison with Traditional Latent Variance

We further compare our state-derived explicit regulation with traditional latent variance which employs learnable Gaussian parameterization to model probabilistic adaptation (Koleilat et al., 2026). Both methods use the same backbone model, adapter architecture and segmentation head for fair comparison. As shown in Fig. 6(a), deriving modulation from proposed pixel-text state removes independent variance heads, which brings a reduction of 14.10% trainable parameters. In Fig. 6(b)-(e), this smaller parameterization can even yield slightly higher observed mean DSC across all four modalities. This results support that our proposed explicit regulation offers a favorable trade-off between parameter cost and segmentation performance under limited supervision.

B.2 Analysis of Selective Spatial Rendering

To verify the function of our SSR, we conduct comprehensive experiments to check whether prediction margins identify useful refinement locations, whether local detail corresponds to those locations, and how the proposed half-resolution stage affects the learned full-resolution corrections.

State-guided Spatial Allocation. Fig. 7(a) and (b) compare state-guided allocation, which selects locations using the current pixel-text margin, with vanilla uniform sampling at identical budgets of 2,048 half-resolution and 4,096 full-resolution points. Here we employ boundary coverage to measure the fraction of GT boundary-band pixels included in the full-resolution selection. State-guided allocation covers 64.6–94.7% of these pixels, compared with 8.1–8.5% for uniform sampling, and improves boundary F1 by 2.48–5.09 percentage points across the four modalities. The coverage and accuracy gains together show that the current prediction state provides a useful spatial criterion for allocating the refinement budget.

Spatial Correspondence of Visual Detail. Fig. 7(c) and (d) keep the native update locations fixed and shuffle detail vectors among points with the same margin sign and similar margin magnitudes, which disrupts spatial correspondence while retaining the available detail vectors. Using the correctly matched detail yields higher boundary F1 and DSC in all four modalities, which indicates that SSR uses location-specific visual information at the state-selected points to improve its corrections.

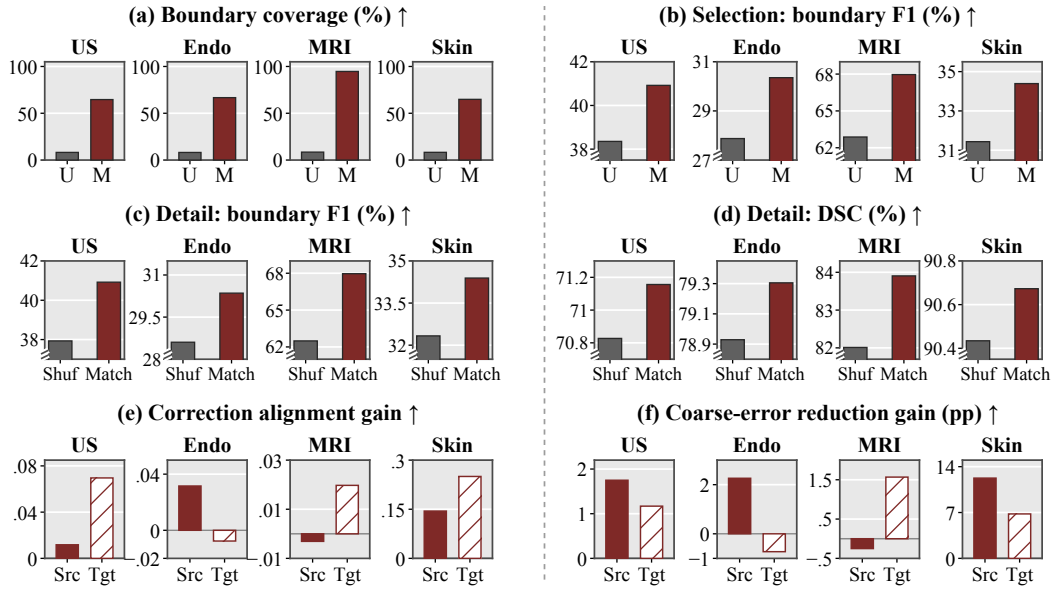

Figure 7: Analysis of selective spatial rendering. Boundary coverage measures the fraction of GT boundary-band pixels selected for refinement. Correction alignment measures the normalized agreement between full-resolution probability updates and the correction needed to match GT. Coarse-error reduction measures the fraction of coarse squared probability error removed by these updates. Panels (e,f) report gains over the variant trained without the half-resolution stage.

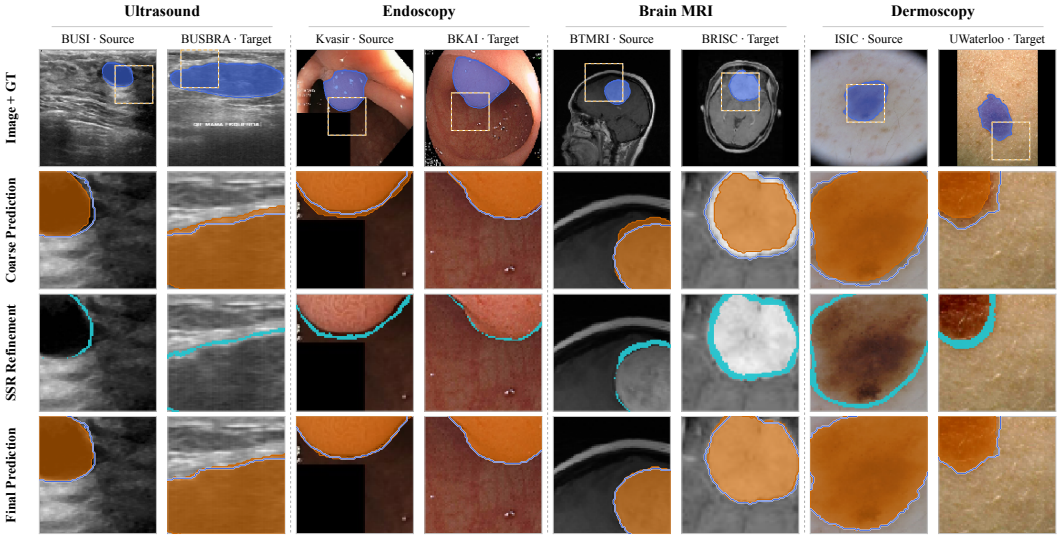

Figure 8: Qualitative visualization of SSR. Blue and orange denote GT annotations and segmentation predictions, respectively. Yellow dashed boxes mark the regions enlarged below, where blue contours indicate GT boundaries. Cyan highlights pixels whose predicted labels change from the coarse to the final prediction.

Effect of the Half-Resolution Stage. Fig. 7(e) and (f) examine full-resolution corrections learned with and without half-resolution rendering. Both models are queried at the union of their selected full-resolution locations, and each correction is evaluated against its own coarse prediction. Let p_0 denote the coarse probability, p_1 the probability after the queried full-resolution correction, and y

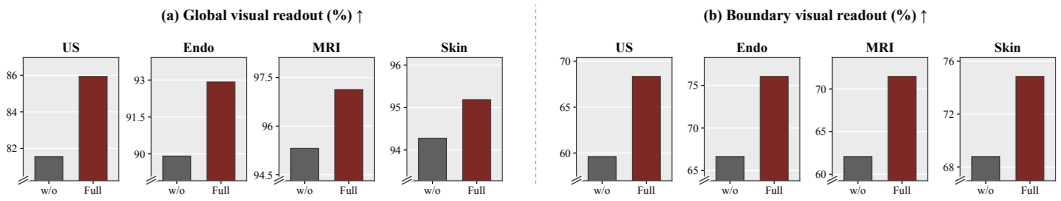

Figure 9: Analysis of text-to-vision attention. w/o denotes the variant without T→V attention, while Full denotes our complete implementation. Global visual readout denotes balanced accuracy from source-trained probes over the full image. Boundary visual readout is calculated by averaging foreground and background recall within eight pixels of the GT boundary.

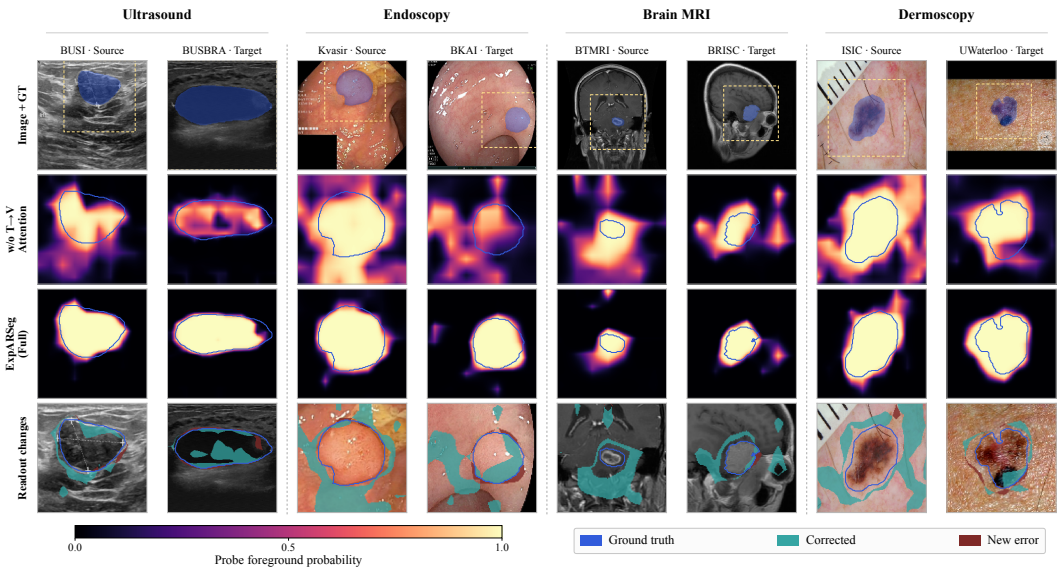

Figure 10: Qualitative visualization of text-to-vision attention. Blue marks GT regions, and yellow dashed boxes indicate the crops shared by subsequent rows. At a 0.5 threshold, the bottom row marks errors corrected by Full relative to w/o in cyan and newly introduced errors in red.

the GT. With $r = y - p_0$ and $d = p_1 - p_0$, we measure

$$C = \frac{\mathbb{E}[rd]}{\sqrt{\mathbb{E}[r^2]\mathbb{E}[d^2]}}, \qquad \eta = \frac{\mathbb{E}[r^2] - \mathbb{E}[(r-d)^2]}{\mathbb{E}[r^2]}. \tag{27}$$

Here, C measures whether probability updates follow the correction needed to match GT, while η measures the fraction of coarse squared error removed. Moments are averaged over image contents within each dataset before computing these ratios. Fig. 7(e) and (f) report full minus w/o-half differences. Both differences are positive in most modality-domain comparisons, with the largest gains in dermoscopy. This supports a role for half-resolution training in shaping error-aligned full-resolution corrections, with effects that vary across modalities and domains.

Visual Presentation. We present several examples to illustrate how SSR refines coarse predictions. As shown in Fig. 8, the SSR refinement can both recover missed lesion region and reduce over-segmentation near lesion boundaries, which brings the final prediction into closer agreement with the GT region to yield a more accurate delineation of lesion extent.

B.3 Analysis of State-derived Bidirectional Attention

We further analyze our bidirectional attention design to verify that both interaction directions serve complementary roles rather than merely adding architectural complexity.

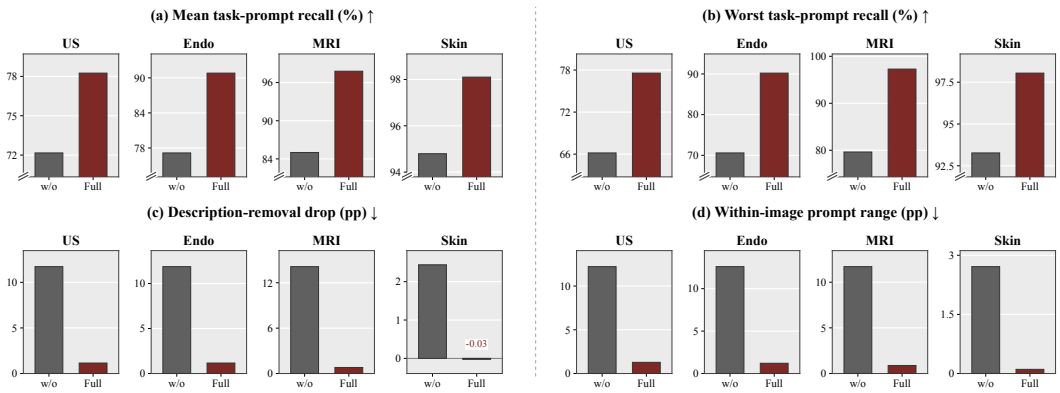

Figure 11: Analysis of vision-to-text attention. w/o denotes the variant without V→T attention, while Full denotes our complete implementation. Mean task-prompt recall averages the fraction of GT foreground retained across six generic prompt templates, using offline thresholds that maximize recall while limiting the background false-positive rate to 5%. Worst task-prompt recall measures the per-image minimum across these prompt templates. Description-removal drop is original-description recall minus mean task-prompt recall. Within-image prompt range denotes the difference between the highest and lowest recall across the six templates.

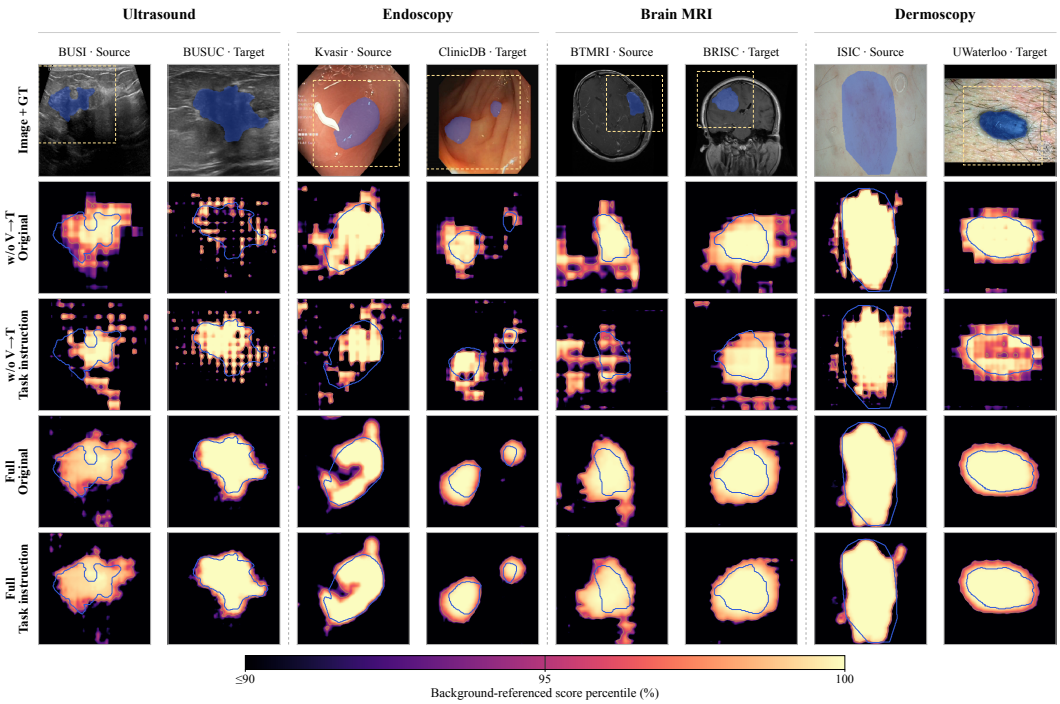

Figure 12: Qualitative visualization of vision-to-text attention. Original denotes the official text prompts used in all our experiments, while Task instruction denotes a fixed prompt template shared across images of the same modality, such as "Segment the lesion in this breast ultrasound image.". Heatmaps report background-referenced score percentiles, which measure the proportion of real background pixels in the image that a given pixel outranks in foreground score.

Increased Visual Discrimination. We fit linear probes on source training data to compare the frozen visual features of Full and the variant without T→V attention, both trained with 10% annotations. As shown in Fig. 9, Full improves global balanced accuracy by 0.91–4.38 percentage points, with larger gains of 6.05–9.44 points near GT boundaries. We also visualize the probe maps in Fig. 10

Table 4: Analysis of the target-conditioned KL regularizer. Opposing gradient fraction measures the proportion of absolute gradient magnitude that reinforces evidence conflicting with GT label. Opposing gradient retained measures this opposing magnitude relative to unconditional KL which penalizes evidence for all classes. Output-GT stop-gradient blocks backpropagation through evidence for the correct class in the KL loss on the final prediction, while leaving the loss value unchanged.

Metric	Method	US	Endo	MRI	Skin
Opposing gradient fraction (%) ↓	Unconditional KL	10.866	2.560	9.419	2.037
	Output-GT stop-gradient	0.368	0.194	0.026	0.100
	TC-KL (ours)	**0.083**	**0.019**	**<0.001**	**0.004**
Opposing gradient retained (%) ↓	Unconditional KL	100.00	100.00	100.00	100.00
	Output-GT stop-gradient	2.49	6.81	0.25	4.68
	TC-KL (ours)	**0.56**	**0.65**	**<0.001**	**0.20**

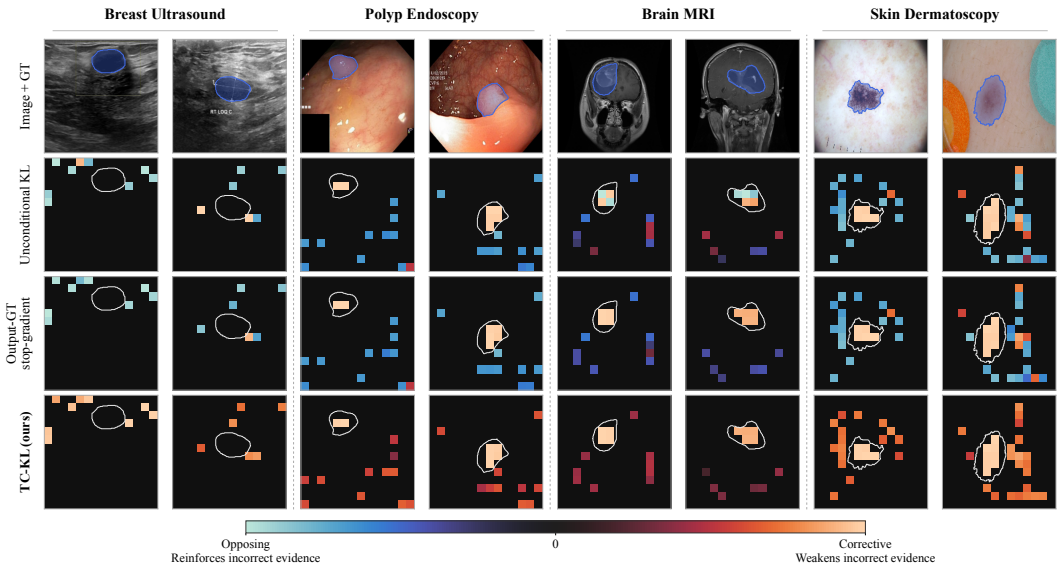

Figure 13: Qualitative visualization of the target-conditioned KL regularizer. Blue masks and light-gray contours denote ground truth. Heatmaps visualize KL feedback where cool colors indicate gradients reinforcing incorrect evidence and warm colors indicate gradients weakening it. TC-KL reduces opposing feedback that persists after output gradient blocking, which manifests as shifting the corresponding regions from cool to warm colors.

to show how these differences appear spatially, where Full recovers more foreground regions and suppresses background responses in the illustrated source and target cases. Together, these results indicate that the text-to-vision interaction makes lesion-background distinctions more accessible in visual features, particularly around boundaries.

Reduced Prompt Dependence. We assess V→T attention by replacing the official segmentation prompts used in all our experiments with generic templates specifying only the segmentation target and modality. Fig. 11 has shown that using native segmentation scores under an offline background false-positive constraint of 5%, Full improves mean foreground recall over the variant without V→T attention by 3.31–13.66 percentage points across the four modalities, while also achieving higher worst-instruction recall. The mean recall drop after replacing the original prompts is at most 1.17 points for Full, compared with 2.44–14.13 points for the variant. Fig. 12 also provides a spatial view of this behavior, showing that Full retains high foreground rankings within lesion regions under generic instructions in source and target examples. Together, these results indicate that incorporating visual information into the text stream reduces reliance on specific segmentation prompts.

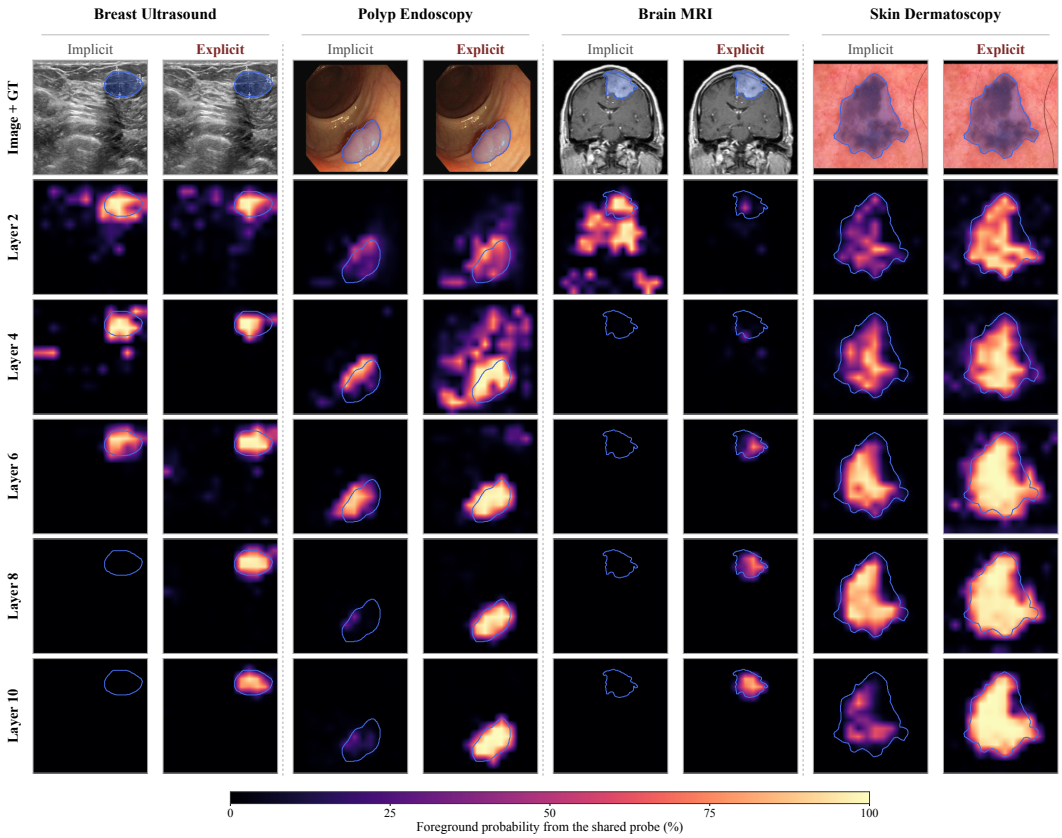

Figure 14: Layer-wise Visualization of Lesion Discrimination.

B.4 Analysis of Target-conditioned KL Regularizer

We analyze how the auxiliary TC-KL changes the regularization feedback received by EPT states. To assess the role of label conditioning, we first compare TC-KL with an unconditional KL variant that penalizes evidence for all classes. We further compare with another variant named Output-GT stop-gradient, which preserves the unconditional KL loss value but blocks backpropagation through evidence for the correct class in the KL term on the final prediction, to test whether penalizing evidence for the correct class contributes to feedback that reinforces incorrect intermediate evidence. As shown in Tab. 4, blocking this gradient path leaves only 0.25–6.81% of the opposing gradient magnitude observed under unconditional KL, identifying it as a substantial source of opposing feedback. TC-KL reduces this magnitude further in each modality, retaining at most 0.65% of the unconditional baseline and lowering the opposing gradient fraction to at most 0.083%. This distinction is also visible in Fig. 13, where cool regions persisting after output gradient blocking turn warm under TC-KL, indicating that the corresponding gradients shift from reinforcing incorrect evidence to weakening it. The agreement between these local changes and the quantitative reductions clarifies that TC-KL uses label conditioning to suppress regularization feedback reinforcing incorrect evidence during adaptation, beyond what output gradient blocking alone achieves.

B.5 Qualitative Comparison between Explicit and Implicit Mechanism

Fig. 14 visualizes lesion readouts at evenly spaced adapter layers across four modalities. In deeper layers, explicit adaptation produces stronger and more complete lesion responses, whereas implicit responses weaken or remain incomplete. These patterns complement Fig. 5, illustrating how explicit regulation helps preserve lesion information in deeper representations.

B.6 Additional Segmentation Visualization

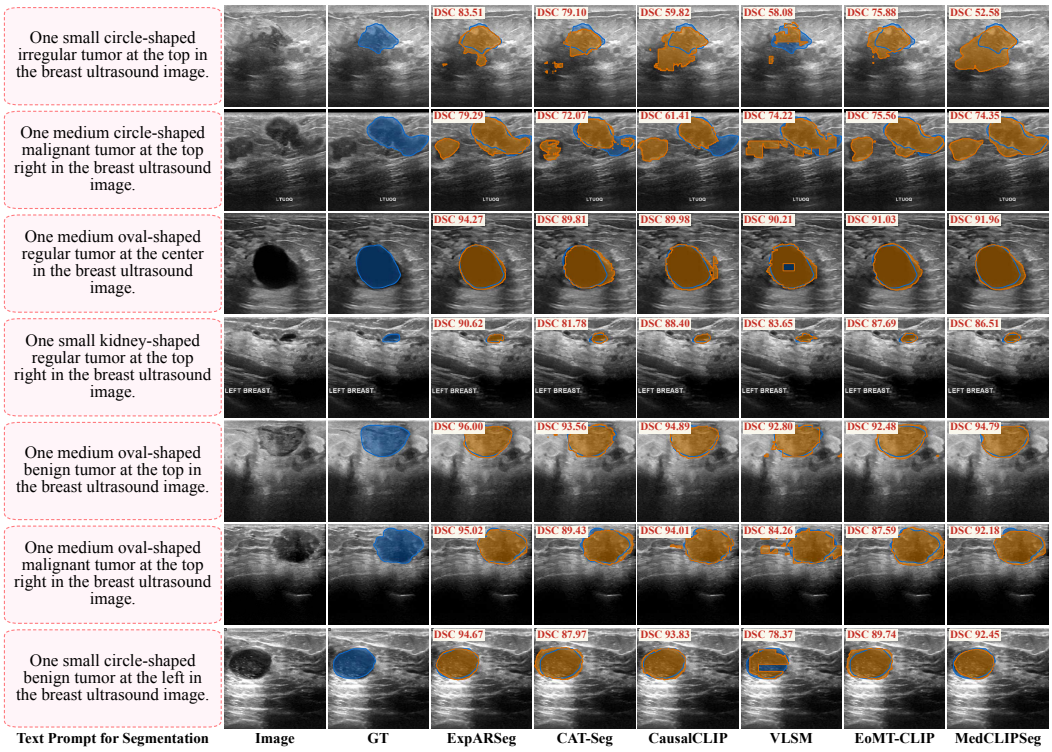

Figure 15: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on BUSI.

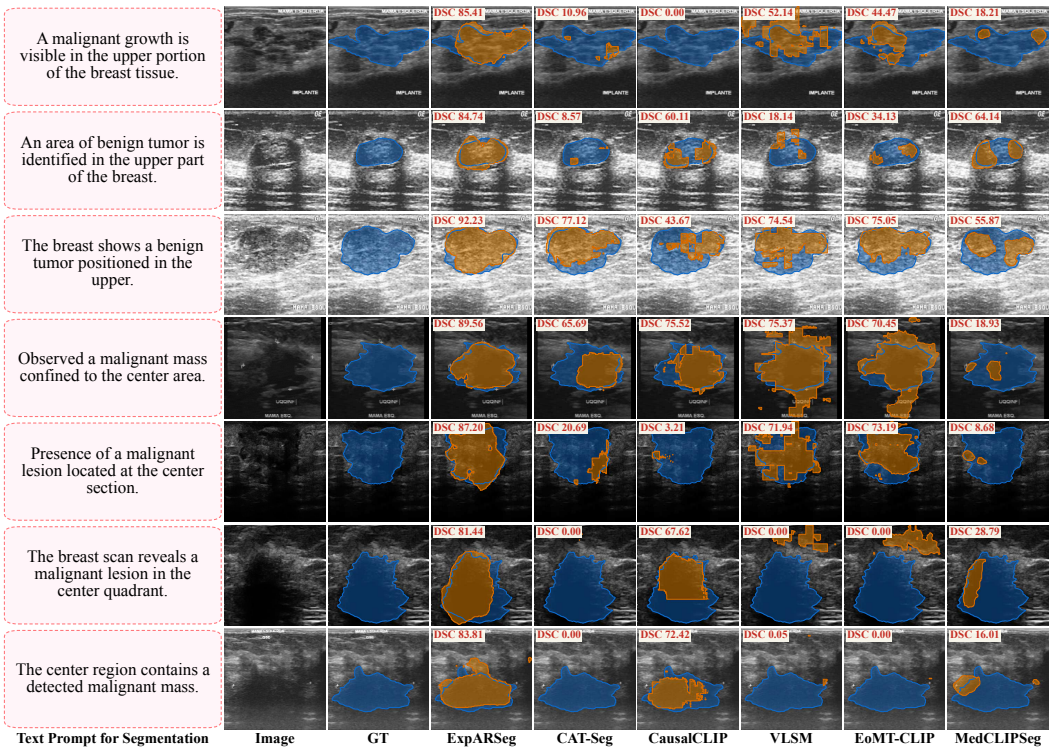

Figure 16: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on BUSBRA.

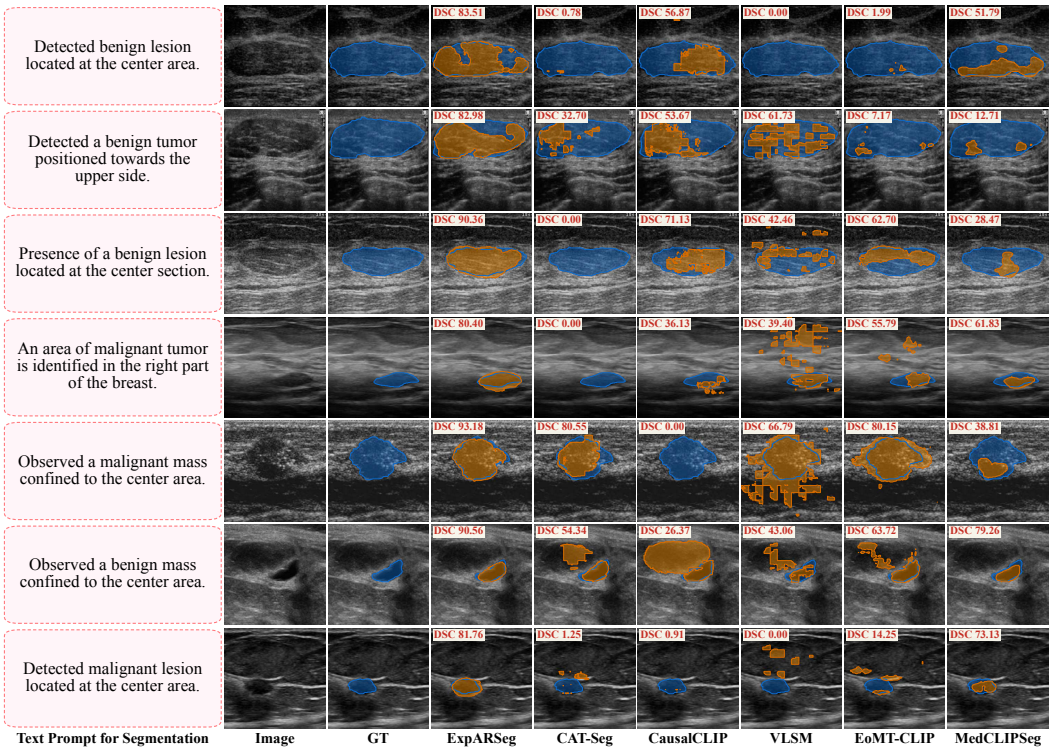

Figure 17: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on BUSUC.

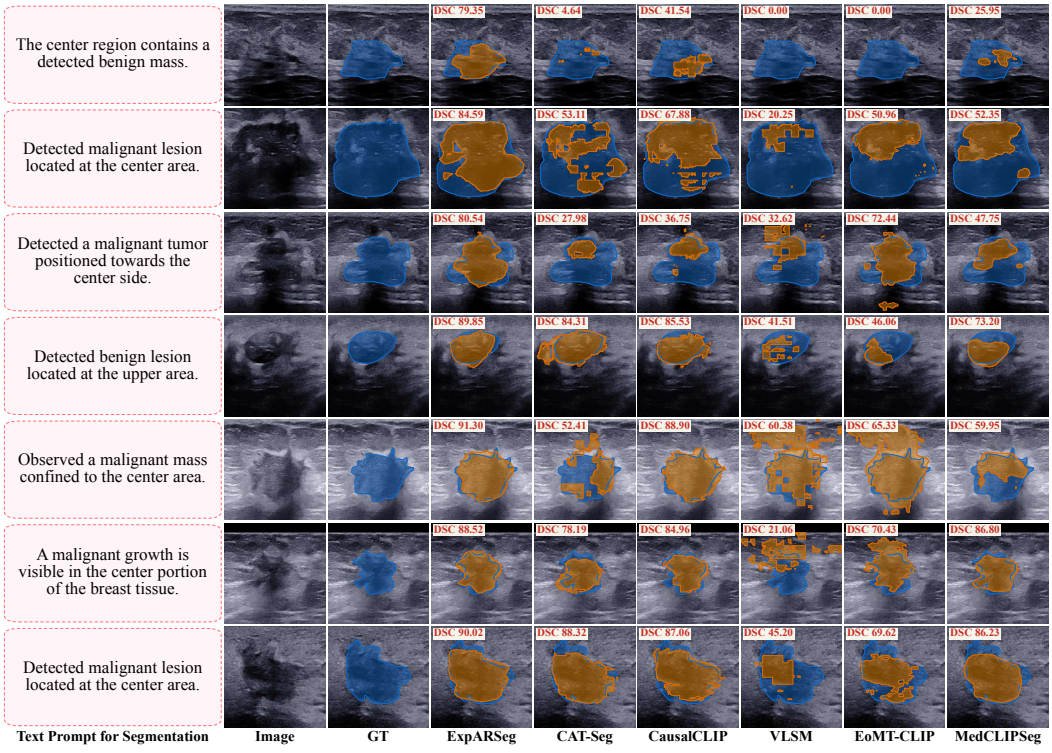

Figure 18: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on BUID.

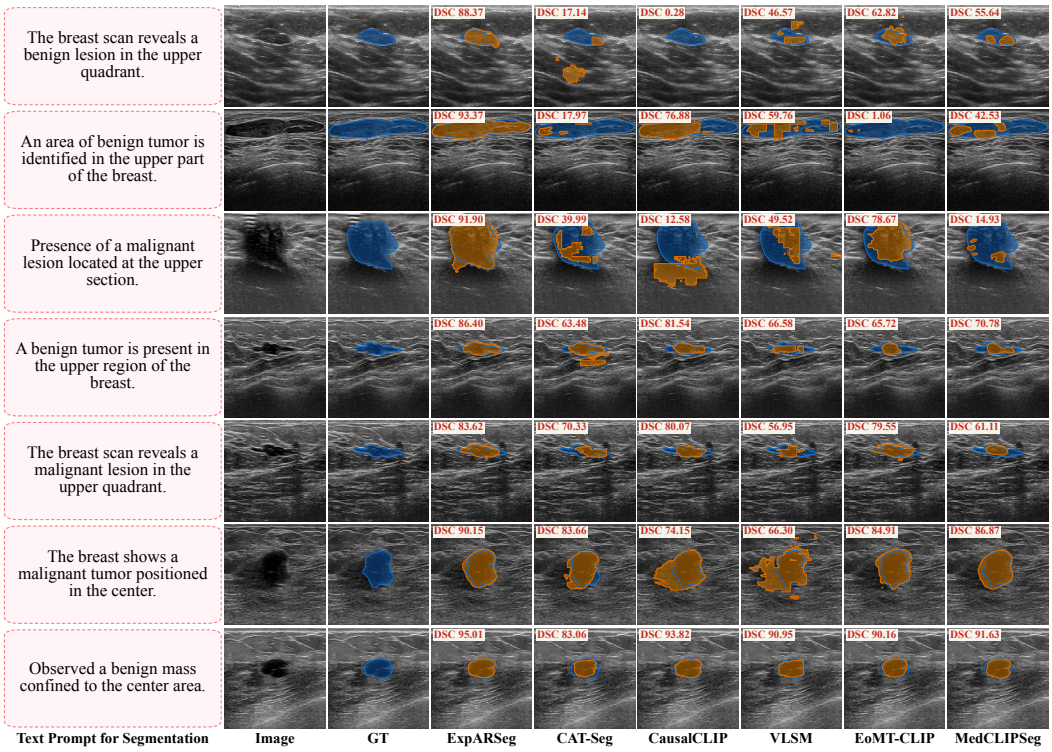

Figure 19: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on UDIAT.



Figure 20: Segmentation visualization of ExpARSeg and state-of-the-art baselines on Kvasir-SEG.

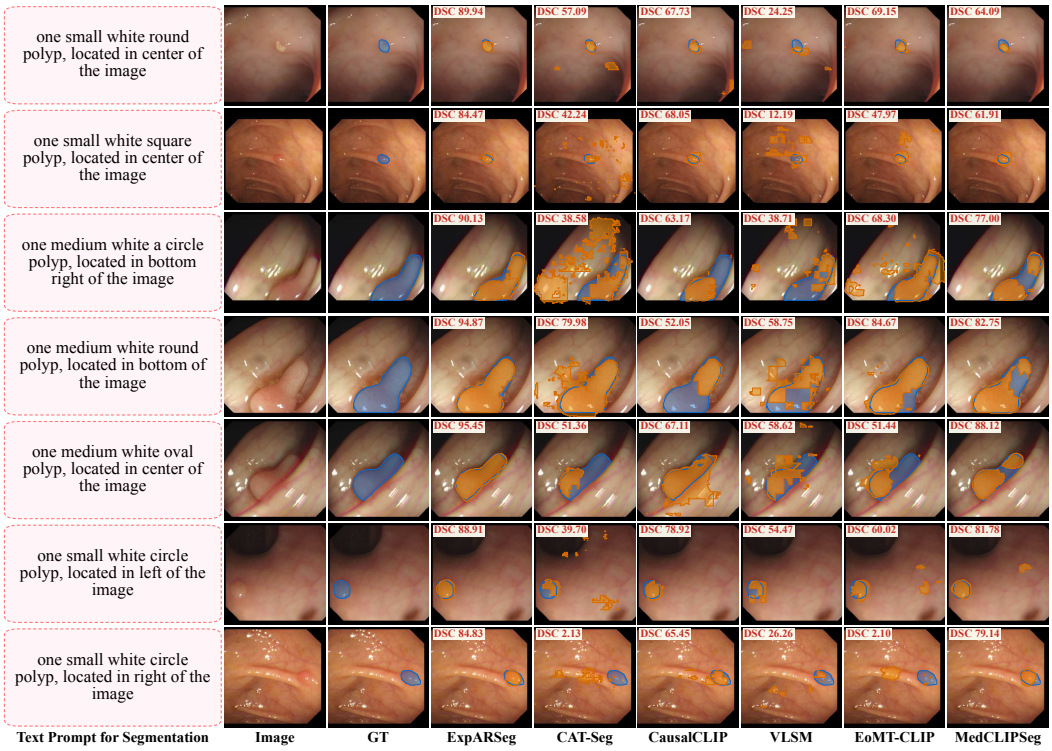

Figure 21: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on CVC-ColonDB.

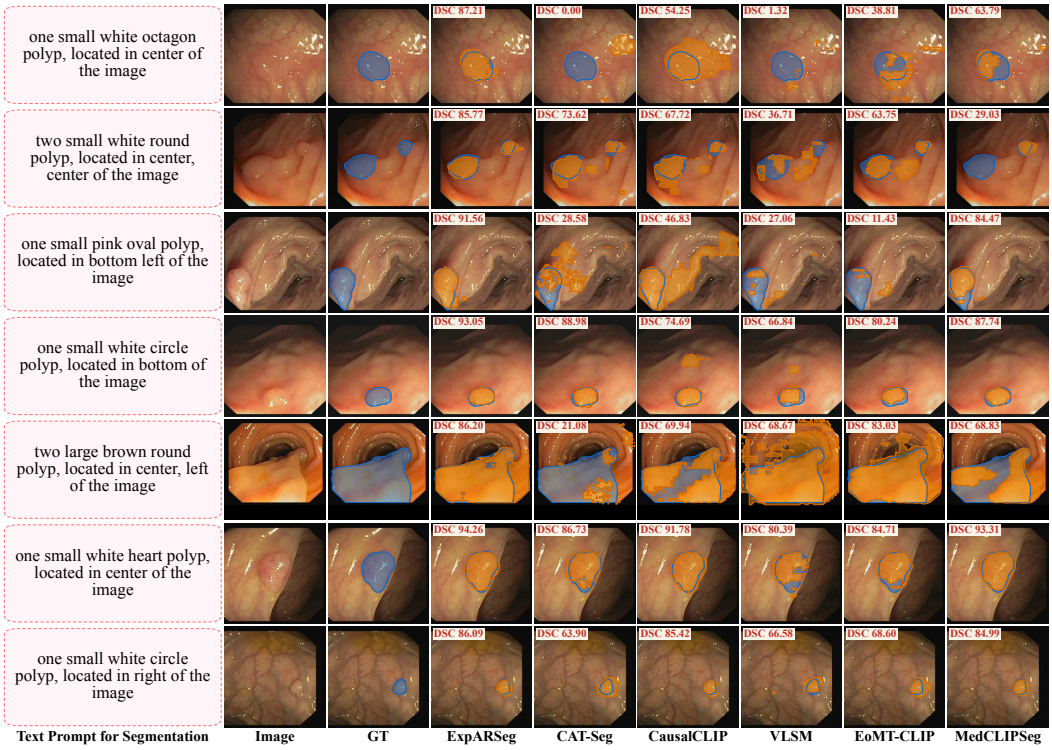

Figure 22: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on CVC-ClinicDB.

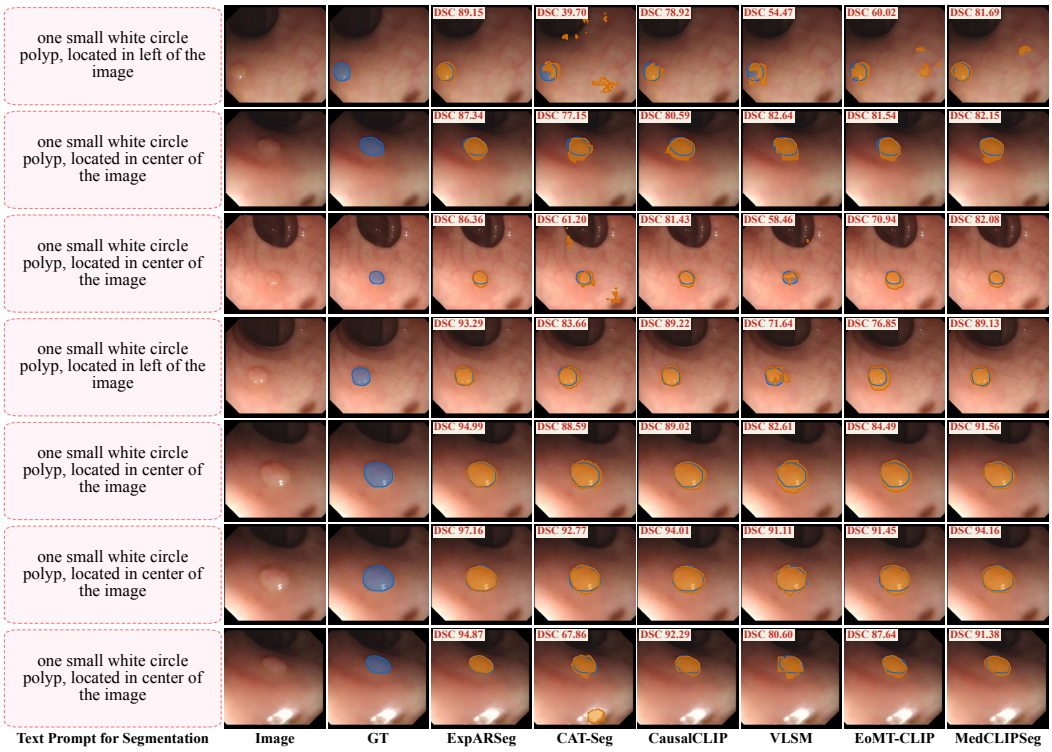

Figure 23: Segmentation visualization of ExpARSeg and state-of-the-art baselines on CVC-300.



Figure 24: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on BKAI.

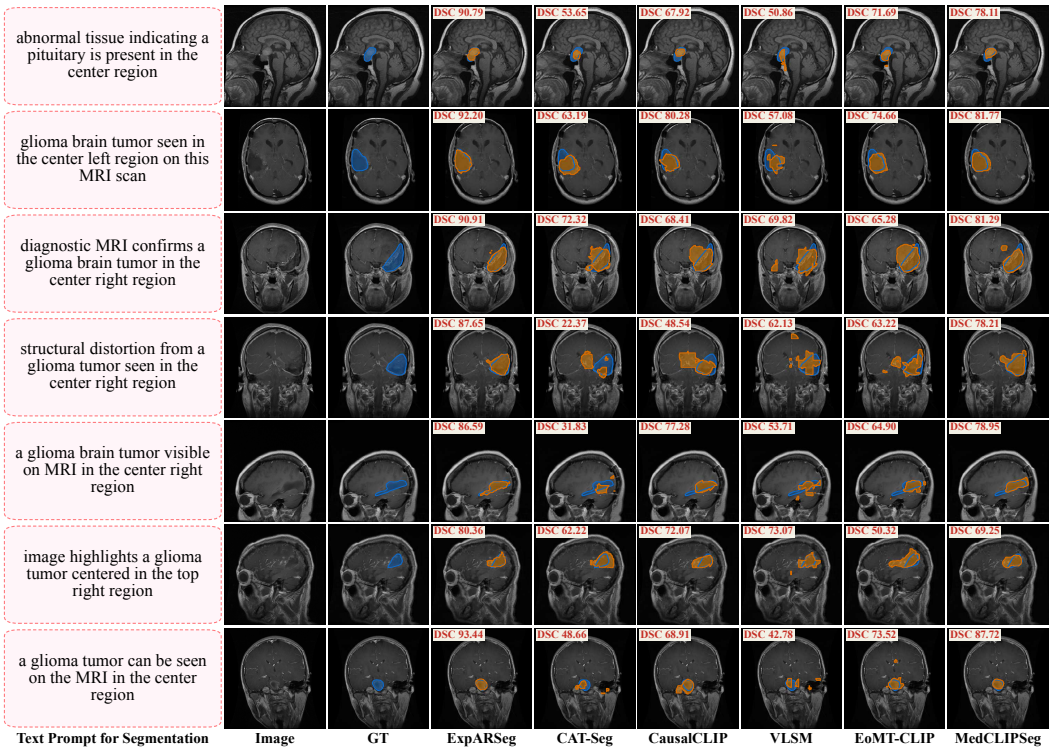

Figure 25: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on BTMRI.

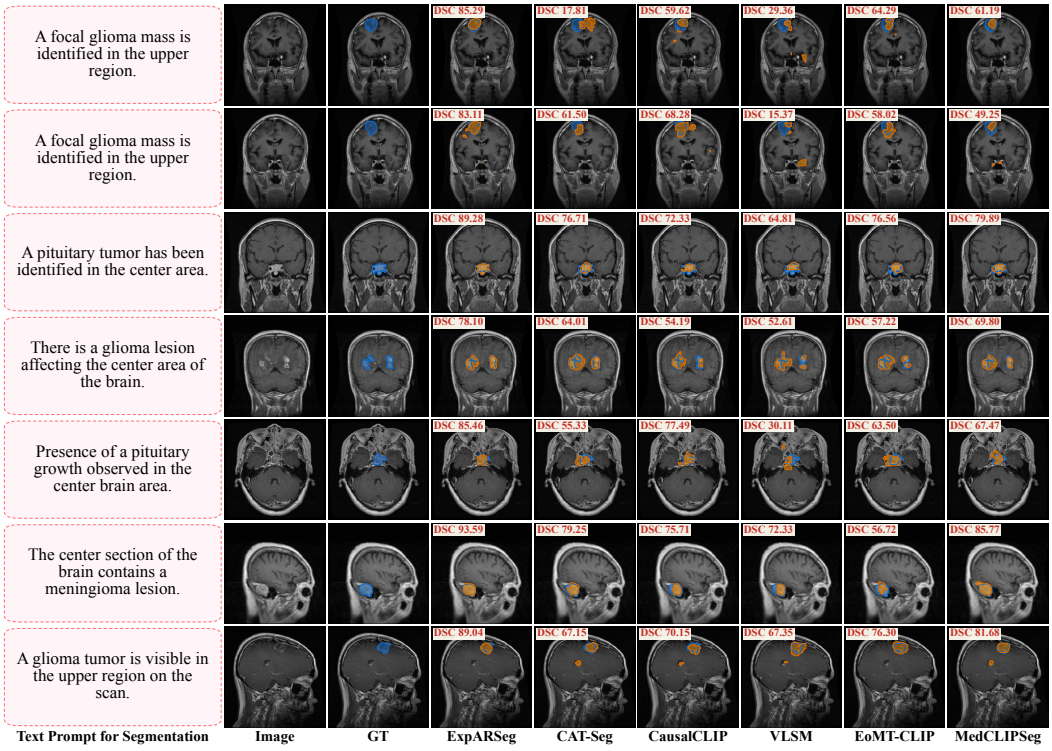

Figure 26: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on BRISC.

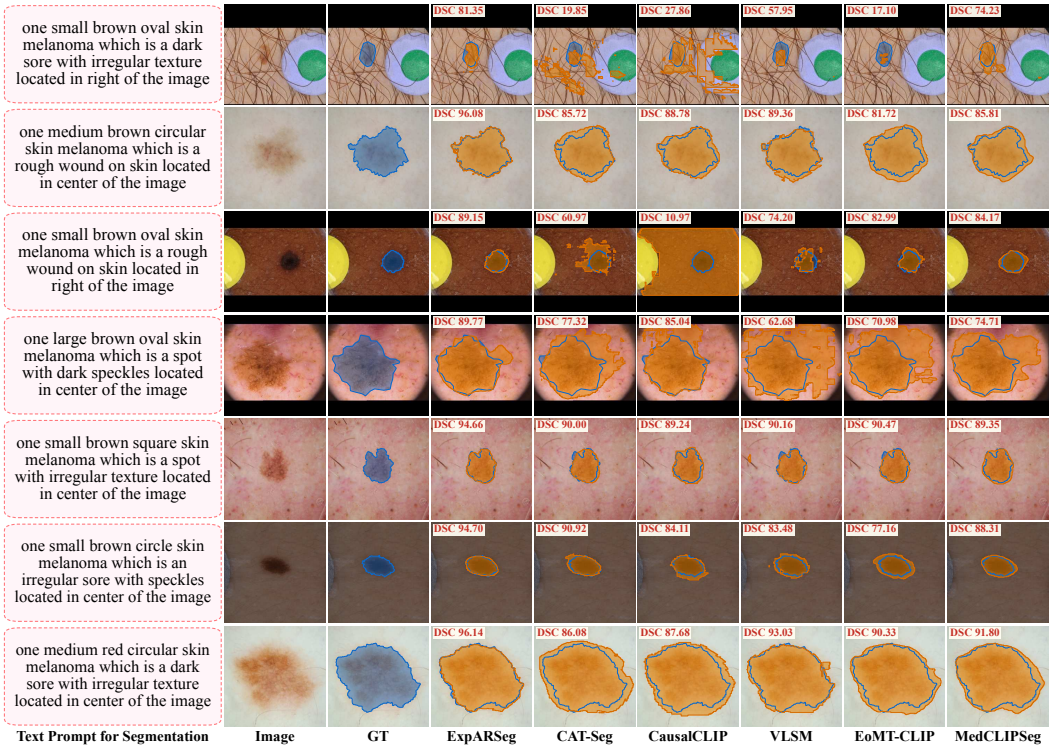

Figure 27: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on ISIC.

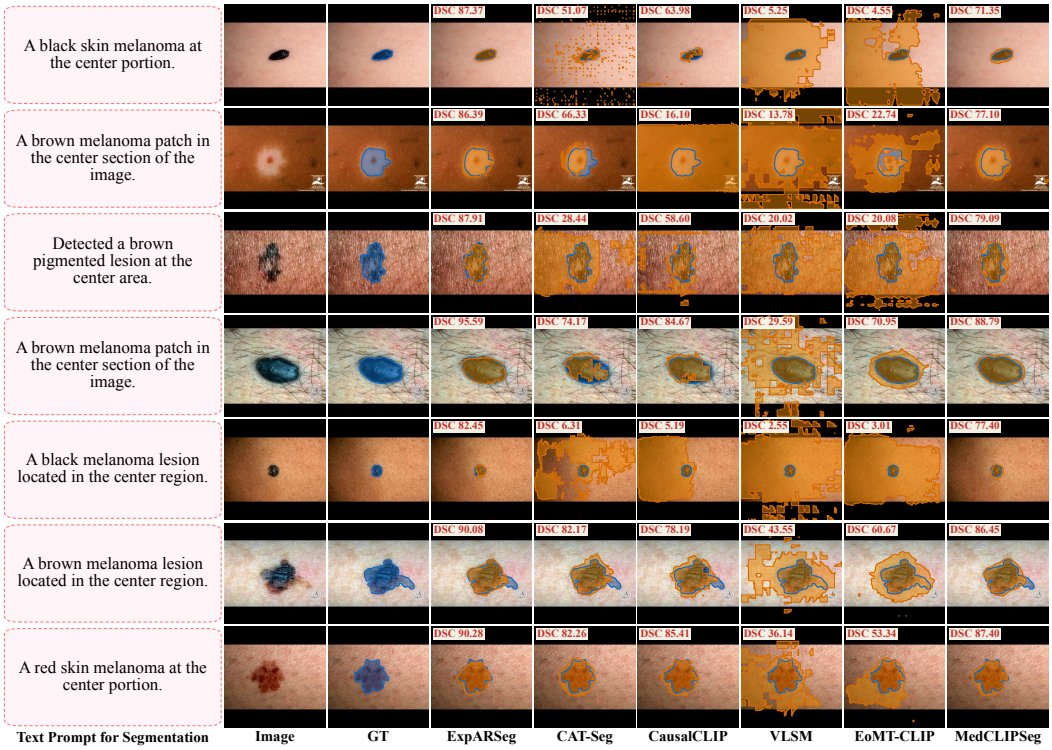

Figure 28: Segmentation visualization of our ExpARSeg and state-of-the-art baselines on UWaterlooSkinCancer.