S^3-ADAPT: GROUNDED SEMANTIC TRANSFER BETWEEN FOUNDATION MODELS FOR MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In medical segmentation cascades built around independently encoded image–text similarity and decoder prompts, language can guide mask prediction without shaping the segmentation encoder's spatial representations. We introduce S^3-Adapt, a framework that connects a pre-aligned vision–language model with a promptable segmentation foundation model through Spatial Grounding, Semantic Transfer, and Selective Correction. Reciprocal Semantic Grounding integrates query-local visual context and reciprocal region–token associations into the source visual encoder, with dense grounding supervision shaping query-conditioned spatial features. Hierarchical Cross-Foundation Semantic Transfer projects these features into multiple stages of the segmentation encoder and jointly conditions gated residual updates on source semantics and the current target state. Point, continuous-mask, and text prompts complement this representation-level transfer to form a reference prediction. Evidence-Conditioned Selective Correction then combines local image evidence with grounding and prediction context to produce a single signed correction gate. The gate controls the direction and bounded magnitude of a logit update while the reference pathway remains fixed. The original parameters of both pretrained encoder backbones remain frozen. Experiments on four public medical image segmentation datasets demonstrate the effectiveness of S^3-Adapt and competitive segmentation performance with only 1.97M task-adapted parameters. Source code is provided in the supplementary material.

1 INTRODUCTION

Medical image segmentation delineates the spatial extent of anatomical structures and lesions, providing a basis for quantitative imaging. U-Net and Transformer-based models recover local detail and global structure (Ronneberger et al., 2015; Chen et al., 2024), but must learn target appearance and location from image–annotation pairs. Language-guided methods instead bring category, region, and contextual descriptions into spatial reasoning. LViT integrates text during representation learning, RecLMIS learns fine-grained correspondences through conditional interaction and cross-modal reconstruction, and HiMix conditions different decoding stages on hierarchical text representations (Li et al., 2024; Huang et al., 2025; Hwang et al., 2026). Although effective, such task-specific fusion often requires elaborate interaction modules.

Parameter-efficient adaptation avoids redesigning full fusion networks or updating large backbones by concentrating task learning in a small parameter subset, thereby reducing the task-specific weights and optimizer states stored separately (Hu et al., 2022; Chen et al., 2022). The central challenge is to convert existing image–text associations, within this limited budget, into semantic conditions that shape spatial representations.

One interface reads regional responses from the shared semantic space of a pre-aligned vision–language model (VLM), as shown in Figure 1(a). CLIP aligns image and text representations through image-level contrastive learning but provides no direct pixel-level boundary supervision (Radford et al., 2021). MaskCLIP obtains dense labels by local image–text matching, while MedCLIPSeg improves fine-grained representations through probabilistic cross-modal adaptation and local contrastive learning (Zhou et al., 2022; Koleilat et al., 2026). These methods strengthen local

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.

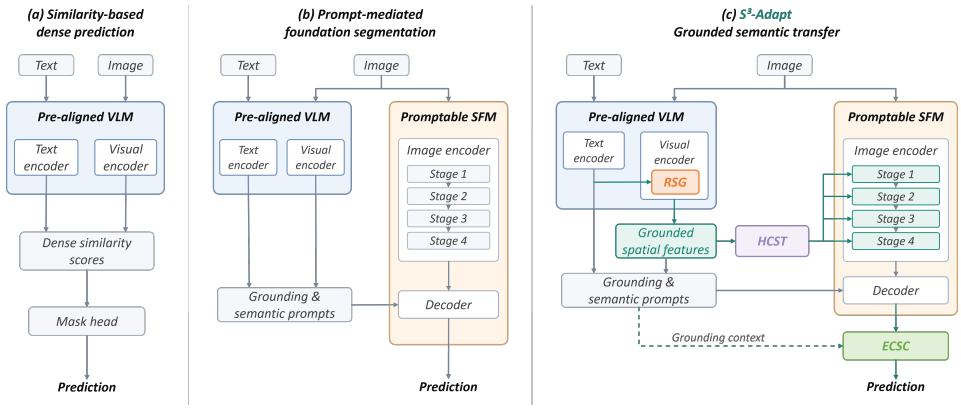

Figure 1: Representative VLM–SFM interfaces: (a) dense similarity readout, (b) grounding-derived prompts, and (c) our grounded representation transfer with HCST and ECSC.

semantic discrimination, yet must convert relevance into shape- and boundary-consistent predictions.

A second interface converts grounding into prompts and uses a promptable segmentation foundation model (SFM) to generate masks (Figure 1(b)), as in the zero-shot pathway of MedCLIP-SAMv2 (Kirillov et al., 2023; Koleilat et al., 2025). When image and text are encoded independently and models communicate mainly through prompts, source interaction occurs after encoding and source spatial representations do not update the target hierarchy. OpenWorldSAM builds multi-instance queries from early vision–language fusion features and derives instance-aware soft prompts through cross-attention with SAM2 encoder outputs (Xiao et al., 2025). Its cross-model interaction, however, remains in prompt construction and updates instance queries rather than target encoder features.

We introduce S^3-Adapt (Spatial Grounding · Semantic Transfer · Selective Correction), which establishes the representation-level pathway in Figure 1(c). Reciprocal Semantic Grounding (RSG) combines query-local context, reciprocal region–token associations, and dense supervision to form query-related representations during source encoding. Hierarchical Cross-Foundation Semantic Transfer (HCST) uses level-specific projections to inject these representations into the target hierarchy and constructs gated, target-state-dependent residuals, enabling semantics to shape subsequent encoder features rather than only prompts. Point, continuous-mask, and text prompts then support a reference prediction. On this fixed reference, Evidence-Conditioned Selective Correction (ECSC) uses a signed gate to bound the direction and magnitude of local logit updates. Both encoder backbones remain frozen; only task-adaptation parameters are updated, allowing foundation weights to be shared while independent adaptation states are maintained across segmentation tasks.

Our main contributions are as follows:

- **Grounding-constrained segmentation.** Source spatial semantics are transferred within target encoder layers instead of being compressed only into decoder prompts.
- **Coordinated RSG and HCST.** Reciprocal interaction and grounding supervision form query-related representations, while source semantics and target states jointly condition hierarchical residual updates.
- **Selective correction.** An image- and prediction-conditioned tanh gate performs bounded correction without updating the reference model.

2 Related Work

2.1 Vision–Language Interaction for Medical Segmentation

Medical segmentation baselines span convolutional, self-configuring, and Transformer architectures. U-Net and UNet++ use skip-connected encoder–decoders, nnU-Net configures its pipeline to

each dataset, and Swin-Unet and TransUNet introduce Transformer-based context modeling (Ronneberger et al., 2015; Zhou et al., 2018; Isensee et al., 2021; Cao et al., 2023; Chen et al., 2024). These designs progressively improve multiscale fusion, robustness, and long-range context, but infer target extent solely from image–annotation pairs.

Language-guided methods differ mainly in where cross-modal interaction occurs. GLoRIA learns global–local image–report alignment, whereas LAVT-RS and LanGuideMedSeg introduce language within segmentation networks (Huang et al., 2021; Yang et al., 2025; Zhong et al., 2023). LViT uses a U-shaped text branch, MMI-UNet adds visual–textual matching to the encoder, RecLMIS learns conditioned correspondences and reconstruction, and MTGT and HiMix organize multiscale or hierarchical text guidance (Li et al., 2024; Bui et al., 2024; Huang et al., 2025; Zhao et al., 2026; Hwang et al., 2026). EviVLM instead uses evidential learning to quantify and reduce the modality gap (Pan et al., 2026). BTHA improves reuse across backbones with shape-preserving multiscale adaptation, gated residuals, and hierarchical supervision (Liu et al., 2026). Although these strategies strengthen task-specific fusion at different encoding or decoding stages, their visual and language pathways remain components of a single segmentation design. S^3-Adapt instead transfers grounded spatial representations from an independent VLM into a separate SFM encoder.

CLIP and BiomedCLIP provide reusable image–text spaces, MaskCLIP reads dense labels through local matching, and MedCLIPSeg learns fine-grained predictions through probabilistic interaction and region-level contrast (Radford et al., 2021; Zhang et al., 2025a; Zhou et al., 2022; Koleilat et al., 2026). Such semantic spaces improve regional discrimination but do not inherently impose target boundaries. RSG likewise forms query-related source representations, but HCST transfers their full vectors into another encoder; only the scalar grounding readout is routed to prompts.

2.2 Semantic Collaboration and Efficient Adaptation of Foundation Models

SAM and SAM 2 expose promptable segmentation interfaces, while MedSAM and MedSAM2 extend them to medical images, volumes, and videos (Kirillov et al., 2023; Ravi et al., 2025; Ma et al., 2024; 2025). VLM–SFM collaboration often constructs prompts or decoder inputs: MedCLIP-SAMv2 uses semantic saliency, EVF-SAM an early-fusion encoder, OpenWorldSAM instance queries, and ESC-Net correlation-derived pseudo-prompts (Koleilat et al., 2025; Zhang et al., 2025b; Xiao et al., 2025; Lee et al., 2025). DIT injects text instructions into SAM, whereas TextSAM-EUS learns text prompts with lightweight SAM adaptation (Huang et al., 2024; Spiegler et al., 2025). These approaches connect semantics primarily through prompting, instruction tuning, or decoder aggregation rather than transferring independent source representations into target encoder blocks. CLIP-Guided SAM further aligns CLIP patch features and text inside SAM adapters, but uses post-encoding matching and additive fusion (Jalilian & Bais, 2026). In contrast, RSG introduces region–token interaction during source encoding, while HCST derives gated residuals from source semantics, target features, and their interaction.

Parameter-efficient adaptation limits task-specific learning through low-dimensional updates or lightweight branches: LoRA adds low-rank increments, AdaptFormer bottleneck branches, and SAM-Adapter visual-prior adapters (Hu et al., 2022; Chen et al., 2022; 2023). We use these as implementation tools.

3 Method

3.1 Overview

As shown in Figure 2, for a batch of B image–text pairs, two preprocessing pathways produce $\mathbf{I}_{\mathrm{vl}} \in \mathbb{R}^{B\times 3\times 224\times 224}$ and $\mathbf{I}_{\mathrm{seg}} \in \mathbb{R}^{B\times 3\times 512\times 512}$. BiomedCLIP and MedSAM2 serve as the source VLM and target SFM, respectively (Zhang et al., 2025a; Ma et al., 2025). The frozen text encoder returns token representations $\mathbf{T} \in \mathbb{R}^{B\times L\times 768}$ and per-sample normalized global representations $\mathbf{t}_{\mathrm{g}} \in \mathbb{R}^{B\times 512}$, where L is the sequence length. We use $[\cdot,\cdot]$ for feature concatenation, $\odot$ for element-wise multiplication, and $\mathcal{R}_h$ for per-sample bilinear interpolation to an $h \times h$ grid.

RSG updates source image-patch representations to form $\mathbf{S} \in \mathbb{R}^{B\times 768\times 14\times 14}$, which HCST projects into the target hierarchy for local-state-dependent updates. In parallel, matching $\mathbf{S}$ with $\mathbf{t}_{\mathrm{g}}$ yields a grounding map $\mathbf{Z}_{\mathrm{g}}$ for point and continuous-mask prompts, while $\mathbf{t}_{\mathrm{g}}$ generates text prompt to-

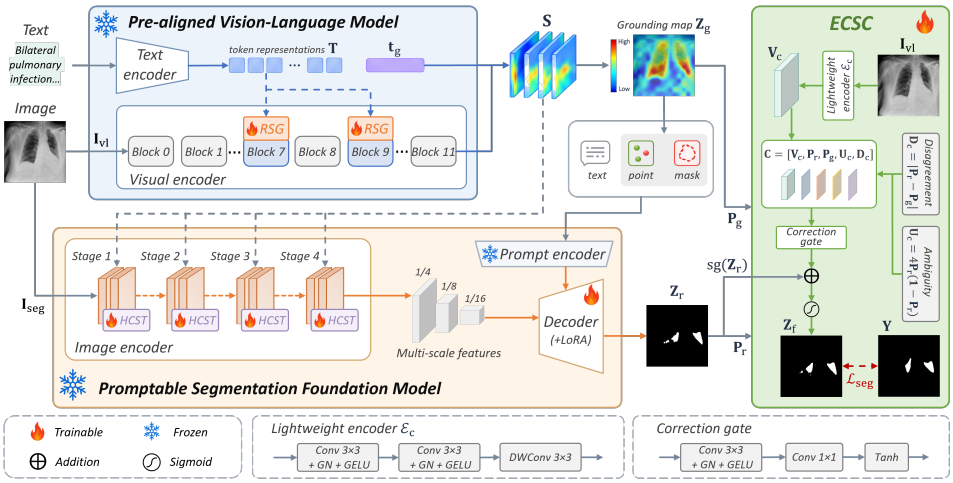

Figure 2: S^3-Adapt overview. RSG forms grounded source representations $\mathbf{S}$, which HCST transfers through four MedSAM2 encoder stages. Grounding and global text provide spatial and text prompts for the reference logits $\mathbf{Z}_\mathrm{r}$; with this pathway frozen, ECSC uses image and prediction evidence to produce the final logits $\mathbf{Z}_\mathrm{f}$.

kens. The adapted native decoder fuses target features and prompts into the reference prediction $\mathbf{Z}_\mathrm{r}$, which ECSC refines into $\mathbf{Z}_\mathrm{f}$ using image and prediction context. All three logit maps belong to $\mathbb{R}^{B\times1\times224\times224}$. Thus, $\mathbf{S}$ is the vector representation transferred across models, whereas $\mathbf{Z}_\mathrm{g}$ is its scalar grounding readout.

3.2 Reciprocal Semantic Grounding

Figure 3 illustrates RSG at a selected source-encoder block. Its local branch preserves neighborhood structure, while reciprocal normalization retains region–token associations supported in both directions before aggregating text semantics.

Query-local context. At selected source block k, let $\mathbf{V}_k \in \mathbb{R}^{B\times N_v\times768}$ denote the image-patch states after the attention residual, where $N_v = 14^2$. Affine projections yield visual and text bottleneck representations $\mathbf{U} = P_v(\mathbf{V}_k)$ and $\mathbf{E} = P_t(\mathbf{T})$, with shapes $B\times N_v\times d$ and $B\times L\times d$; d is the bottleneck width. We reshape $\mathbf{U}$ into a two-dimensional grid, extract context with depthwise convolution, GELU, and pointwise convolution, and add it to $\mathbf{U}$ to obtain queries $\mathbf{Q}$.

Reciprocal region–token associations. For sample b, let $\mathcal{J}_b$ be the valid text-token indices, and let $\mathbf{q}_{bi}$ and $\mathbf{e}_{bj}$ be region and token vectors in $\mathbf{Q}$ and $\mathbf{E}$. Bias-free projections and channel-wise normalization give $\widehat{\mathbf{q}}_{bi} = \mathrm{nrm}(W_q\mathbf{q}_{bi})$ and $\widehat{\mathbf{e}}_{bj} = \mathrm{nrm}(W_k\mathbf{e}_{bj})$, where nrm denotes ℓ_2 normalization. The region–token score is $A_{bij} = \langle\widehat{\mathbf{q}}_{bi}, \widehat{\mathbf{e}}_{bj}\rangle/\tau$ for $j\in\mathcal{J}_b$, with temperature τ. Normalizing this matrix along both token and region dimensions gives the reciprocal associations:

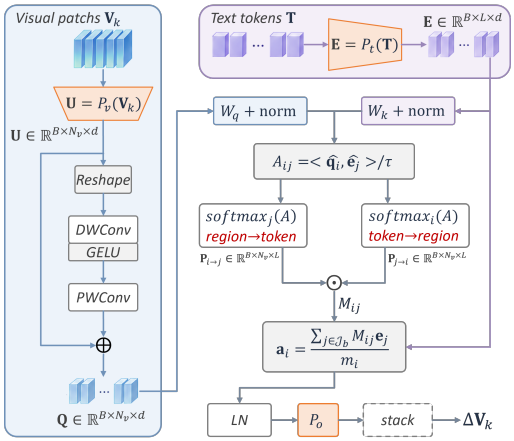

Figure 3: RSG at a selected source-encoder block. Query-local features and reciprocal region–token associations produce a semantic residual for the visual state.

$$M_{bij} = \mathrm{softmax}_j(A_{bij})\ \mathrm{softmax}_i(A_{bij}). \tag{1}$$

The two factors capture a region's relative association with tokens and a token's support over regions, normalized over $j\in\mathcal{J}_b$ and all N_v regions, respectively. Let $m_{bi} = \sum_{j\in\mathcal{J}_b} M_{bij}$. Token semantics

are then aggregated for each visual region:

$$\mathbf{a}_{bi} = \frac{\sum_{j \in \mathcal{J}_b} M_{bij} \mathbf{e}_{bj}}{m_{bi}}. \tag{2}$$

Channel-wise layer normalization and output projection give $\Delta \mathbf{v}_{bi} = P_o\big(\mathrm{LN}(\mathbf{a}_{bi})\big)$, stacked as $\Delta \mathbf{V}_k \in \mathbb{R}^{B \times N_v \times 768}$. This update runs in parallel with the native MLP residual: $\mathbf{V}_k^{\mathrm{out}} = \mathbf{V}_k + \mathcal{F}_k^{\mathrm{vl}}(\mathbf{V}_k) + \alpha_k \Delta \mathbf{V}_k$. Here, P_o restores the channel width, $\mathcal{F}_k^{\mathrm{vl}}$ is the native MLP branch, and α_k is a block-specific learnable signed scale.

Dense grounding supervision. The frozen visual projection W_{proj} maps the final spatial representations $\mathbf{S}$ into the shared image–text space. Let $\mathbf{s}_{bi}$ be the source vector at location i in sample b, and $\mathbf{t}_b$ its global text vector. The native grounding logits are

$$\mathbf{Z}_{\mathrm{g}}^{14}(b,i) = \gamma \langle \mathrm{nrm}(W_{\mathrm{proj}} \mathbf{s}_{bi}), \mathbf{t}_b \rangle. \tag{3}$$

Here, γ is the frozen similarity scale and the superscript 14 denotes the native grid. The supervision-resolution map is $\mathbf{Z}_{\mathrm{g}} = \mathcal{R}_{224}(\mathbf{Z}_{\mathrm{g}}^{14})$. Its loss directly constrains RSG representations through the frozen projection, while the full vectors $\mathbf{S}$ remain available to HCST. Appendix A.1 specifies adaptation locations and gradient boundaries.

3.3 Hierarchical Cross-Foundation Semantic Transfer

Figure 4 shows one HCST unit. The transfer candidate jointly encodes projected source semantics, the current target state, and their element-wise interaction; a target-dependent gate controls the update at each encoder level.

Hierarchical projection and target-state interaction. HCST injects source spatial semantics according to the current segmentation state. We select Hiera blocks 0, 2, 9, and 11 at the ends of the four encoding stages, inserting each update after the attention residual and before the native MLP. These stage-end locations aggregate their respective features while propagating transferred information into subsequent representations. All units read the same final source representation rather than features from different source layers; Table A.1 lists their locations and dimensions.

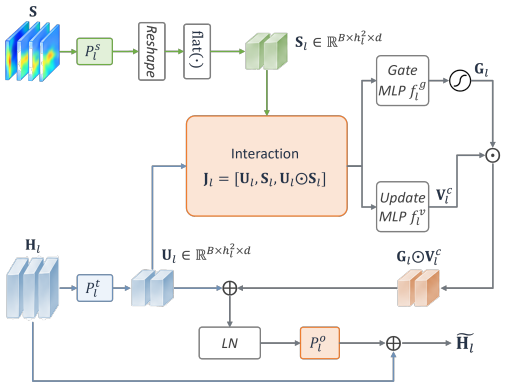

Figure 4: HCST at target-encoder level ℓ. Projected source semantics interact with the current target state to form a gated residual update.

For level $\ell \in \{1,2,3,4\}$, we flatten the target state to $\mathbf{H}_\ell \in \mathbb{R}^{B \times h_\ell^2 \times c_\ell}$, where h_ℓ and c_ℓ are its spatial side length and channel width. A level-specific projection produces the target bottleneck representation $\mathbf{U}_\ell = P_\ell^t(\mathbf{H}_\ell)$. Pointwise projection, grid adaptation, and flattening of the source representation yield $\mathbf{S}_\ell = \mathrm{flat}\big(\mathcal{R}_{h_\ell}(P_\ell^s(\mathbf{S}))\big)$. Both belong to $\mathbb{R}^{B \times h_\ell^2 \times d}$. The interaction features $\mathbf{J}_\ell = [\mathbf{U}_\ell, \mathbf{S}_\ell, \mathbf{U}_\ell \odot \mathbf{S}_\ell]$ retain target content, source semantics, and channel-wise correspondence.

Gated residual transfer. Two independent MLPs read $\mathbf{J}_\ell$ to produce the gate $\mathbf{G}_\ell = \sigma(f_\ell^g(\mathbf{J}_\ell))$ and candidate $\mathbf{V}_\ell^c = f_\ell^v(\mathbf{J}_\ell)$, respectively, where σ is the sigmoid function. The complete encoder update is

$$\widetilde{\mathbf{H}}_\ell = \mathbf{H}_\ell + P_\ell^o(\mathrm{LN}(\mathbf{U}_\ell + \mathbf{G}_\ell \odot \mathbf{V}_\ell^c)). \tag{4}$$

The updated state enters the native MLP as $\mathbf{H}_\ell^{\mathrm{out}} = \widetilde{\mathbf{H}}_\ell + \mathcal{F}_\ell^{\mathrm{seg}}(\widetilde{\mathbf{H}}_\ell)$. Here, P_ℓ^o restores the target width and $\mathcal{F}_\ell^{\mathrm{seg}}$ is the native MLP residual branch. The gate modulates the source–target candidate, while $\mathbf{U}_\ell$ retains a target-side adaptation path. HCST therefore conditions rather than fixes the semantic update, with independent projection and interaction parameters at each level.

Semantic prompts and reference decoding. Alongside representation transfer, the grounding map generates positive and negative points and a continuous-mask prompt, while the global text vector

yields an additional sparse prompt token. Point extraction is deterministic; the continuous mask preserves a differentiable path. The low-resolution feature $\mathbf{F}_{32}$ and high-resolution features $\mathbf{F}_{128}, \mathbf{F}_{64}$ enter the adapted native decoder $\mathcal{D}$:

$$\mathbf{Z}_{\mathrm{r}} = \mathcal{R}_{224}(\mathcal{D}(\mathbf{F}_{32}, \mathbf{F}_{128}, \mathbf{F}_{64}, \mathbf{P}_{\mathrm{s}}, \mathbf{P}_{\mathrm{d}})) . \tag{5}$$

Here, $\mathbf{P}_{\mathrm{s}}$ and $\mathbf{P}_{\mathrm{d}}$ are sparse and dense prompt embeddings with the batch dimension retained. Decoder LoRA adapts selected attention and MLP projections; decoding uses single-mask output and updates selected native output heads. The resulting feature–prompt fusion forms $\mathbf{Z}_{\mathrm{r}}$. Appendix A.2 details prompt generation and parameter coverage.

3.4 Evidence-Conditioned Selective Correction

ECSC is trained after freezing the reference model. A lightweight image encoder $\mathcal{E}_c$ extracts local evidence $\mathbf{V}_c \in \mathbb{R}^{B\times d_c\times 224\times 224}$ from $\mathbf{I}_{\mathrm{vl}}$, which is combined with reference prediction context. The reference and grounding probabilities are $\mathbf{P}_{\mathrm{r}} = \sigma(\mathrm{sg}(\mathbf{Z}_{\mathrm{r}}))$ and $\mathbf{P}_{\mathrm{g}} = \sigma(\mathrm{sg}(\mathbf{Z}_{\mathrm{g}}))$, where sg denotes stop-gradient. The ambiguity map $\mathbf{U}_c = 4\mathbf{P}_{\mathrm{r}} \odot (\mathbf{1} - \mathbf{P}_{\mathrm{r}})$ emphasizes probabilities near 0.5; $\mathbf{1}$ has the same shape as the probability maps. The disagreement map $\mathbf{D}_c = |\mathbf{P}_{\mathrm{r}} - \mathbf{P}_{\mathrm{g}}|$ measures foreground-probability discrepancy between the two pathways. We concatenate these signals as $\mathbf{C} = [\mathbf{V}_c, \mathbf{P}_{\mathrm{r}}, \mathbf{P}_{\mathrm{g}}, \mathbf{U}_c, \mathbf{D}_c]$.

A convolutional fusion function ϕ and a single-channel projection h_c produce the response $\mathbf{A}_c = h_c(\phi(\mathbf{C}))$. Applying tanh yields a signed correction gate $\mathbf{G}_c = \tanh(\mathbf{A}_c)$. The final output is

$$\mathbf{Z}_{\mathrm{f}} = \mathrm{sg}(\mathbf{Z}_{\mathrm{r}}) + \kappa\mathbf{G}_c. \tag{6}$$

Here, $\mathbf{A}_c, \mathbf{G}_c \in \mathbb{R}^{B\times 1\times 224\times 224}$, and $\kappa > 0$ is the logit residual budget.

3.5 Optimization Objectives and Parameter Efficiency

All outputs use $\mathcal{L}_{\mathrm{seg}} = 0.5\mathcal{L}_{\mathrm{BCE}} + 0.5\mathcal{L}_{\mathrm{Dice}}$, with class-balanced BCE and squared soft Dice. The supervision mask is $\mathbf{Y} \in \{0,1\}^{B\times 1\times 224\times 224}$. The first stage jointly learns source representation preparation and reference formation:

$$\mathcal{L}_{\mathrm{ref}} = \mathcal{L}_{\mathrm{seg}}(\mathbf{Z}_{\mathrm{r}}, \mathbf{Y}) + \lambda_{\mathrm{g}}(e)\mathcal{L}_{\mathrm{seg}}(\mathbf{Z}_{\mathrm{g}}, \mathbf{Y}). \tag{7}$$

Here, e is the training epoch and $\lambda_{\mathrm{g}}(e)$ controls grounding supervision. This stage optimizes RSG, HCST, text-prompt projection, Decoder LoRA, and OHA-selected parameters while preserving input gradients through frozen native modules. The second stage fixes all reference states and optimizes only ECSC:

$$\mathcal{L}_{\mathrm{corr}} = \mathcal{L}_{\mathrm{seg}}(\mathbf{Z}_{\mathrm{f}}, \mathbf{Y}). \tag{8}$$

Both pretrained backbones remain frozen, and only adaptation parameters are updated. Appendix A.3 gives the loss definitions and parameter states for both stages.

4 Experiments

4.1 Experimental Settings

Datasets and metrics. We evaluate on four public two-dimensional medical image segmentation datasets: QaTa-COV19 for chest X-rays (Degerli et al., 2022), MosMedData+ for chest CT (Morozov et al., 2020; Li et al., 2024), Kvasir-SEG for endoscopic polyp segmentation (Jha et al., 2020), and ISIC2016 for dermoscopic lesion segmentation (Codella et al., 2018). QaTa-COV19 and MosMedData+ follow the fixed splits and image-specific text annotations provided with LViT (Li et al., 2024); Kvasir-SEG and ISIC2016 use the fixed splits and P6 structured image-specific prompts from MedVLSM (Poudel et al., 2024). Split sizes and text attributes appear in Appendix B.1.

Region overlap is evaluated using per-image mean Dice and foreground intersection-over-union (mIoU). Boundary errors are measured by the 95th-percentile Hausdorff distance (HD95) and average symmetric surface distance (ASSD). All metrics are computed on a 224×224 grid, and distances are reported in pixels. Complete definitions are provided in Appendix B.2.

Implementation details. The implementation uses Python 3.10, PyTorch 2.9.0, and CUDA 12.8 on an Intel Core i9-14900KF CPU and a single NVIDIA GeForce RTX 5090 D GPU. The model is trained in two stages, with the original pretrained encoder parameters kept frozen. The base configuration uses AdamW, a learning rate of 3×10^{-4}, a batch size of 8, a cosine annealing warm-restart scheduler, and a fixed random seed of 42, with deterministic options enabled. All comparison methods are retrained on the same datasets and evaluated using the same metric computation pipeline.

4.2 Comparison with Existing Methods

Table 1: Quantitative comparison on four datasets. Dice/mIoU are percentages, HD95/ASSD are pixels, Params counts trainable parameters, and FLOPs is per image. Best and second-best values throughout the tables are **bolded** and underlined.

Method	Type	Params ↓	FLOPs ↓	QaTa-COV19		MosMedData+		Kvasir-SEG		ISIC2016	
				Dice ↑	mIoU ↑	Dice ↑	mIoU ↑	Dice ↑	mIoU ↑	Dice ↑	mIoU ↑
				HD95 ↓	ASSD ↓	HD95 ↓	ASSD ↓	HD95 ↓	ASSD ↓	HD95 ↓	ASSD ↓
U-Net (Ronneberger et al., 2015)	Vision-only	31.04M	83.75G	77.03	67.36	38.42	29.89	58.22	47.88	85.09	77.07
				29.71	8.60	39.19	16.38	45.79	18.02	20.15	8.37
UNet++ (Zhou et al., 2018)		24.00M	314.47G	65.75	53.46	67.65	54.46	59.78	48.13	89.26	82.29
				49.40	14.17	18.43	4.51	60.73	21.32	23.68	8.04
nnUNet (Isensee et al., 2021)		39.90M	873.57G	79.54	70.09	72.56	59.56	89.23	83.03	88.23	81.16
				28.42	7.77	13.63	3.51	16.69	4.95	24.73	7.87
Swin-Unet (Cao et al., 2023)		27.17M	**12.29G**	78.42	68.96	68.75	55.03	86.24	79.16	89.92	83.04
				27.94	7.75	16.80	4.33	18.79	5.73	12.09	5.08
Trans-UNet (Chen et al., 2024)		105.28M	58.67G	79.48	70.35	65.46	52.01	88.35	82.07	91.22	84.82
				26.32	7.26	17.84	5.08	17.43	4.96	11.19	4.58
GLoRIA (Huang et al., 2021)	Vision–Language	32.52M	16.36G	79.43	70.11	68.52	54.63	86.58	79.99	90.95	84.61
				29.57	8.00	20.13	5.09	23.01	5.94	12.24	4.89
LAVT-RS (Yang et al., 2025)		227.74M	103.69G	83.97	75.46	61.75	47.68	91.75	85.91	92.85	87.16
				13.71	3.85	19.72	5.09	14.16	3.52	9.77	3.67
LanGuideMedSeg (Zhong et al., 2023)		43.95M	24.15G	84.30	75.78	70.64	56.96	91.85	86.85	92.03	86.06
				14.95	4.05	16.87	4.31	13.08	3.33	11.67	4.30
LViT (Li et al., 2024)		39.93M	58.73G	83.34	74.72	71.12	57.02	88.56	81.97	91.44	85.23
				16.07	4.48	15.70	3.97	20.93	5.57	11.66	4.68
MMI-UNet (Bui et al., 2024)		56.18M	23.29G	84.34	75.97	71.33	57.74	91.21	86.04	92.67	86.89
				13.91	3.86	17.00	4.14	12.98	3.30	9.92	3.87
RecLMIS (Huang et al., 2025)		58.93M	56.56G	84.74	76.39	72.66	59.20	88.30	81.69	90.83	84.06
				13.36	3.82	15.81	3.71	16.17	4.62	11.98	4.78
MTGT (Zhao et al., 2026)		56.73M	261.82G	83.21	74.70	71.49	58.14	87.36	80.74	91.43	85.28
				17.97	4.85	13.84	3.63	18.40	5.25	10.78	4.49
HiMix (Hwang et al., 2026)		44.73M	24.38G	84.75	76.29	71.53	57.81	91.26	85.79	92.28	86.42
				14.85	3.93	17.16	4.17	13.56	3.31	10.92	4.15
EviVLM (Pan et al., 2026)		101.49M	84.37G	83.10	74.54	71.41	57.20	84.01	76.26	91.02	84.56
				15.14	4.18	20.11	4.57	24.15	6.82	11.98	4.83
DIT (Huang et al., 2024)	Foundation-Model-based	212.87M	43.36G	84.42	76.02	64.34	50.44	91.84	86.84	92.52	86.68
				12.71	3.76	16.00	4.15	11.50	3.10	9.31	3.79
TextSAM-EUS (Spiegler et al., 2025)		**1.69M**	1083.89G	78.83	68.94	71.79	58.30	90.42	84.64	92.17	86.45
				27.45	7.79	17.10	4.59	16.84	4.83	10.26	4.23
S^3-Adapt (Ours)	Cross-Foundation	1.97M	114.88G	**84.93**	**76.57**	**77.74**	**65.03**	**93.08**	**88.47**	**93.14**	**87.71**
				12.80	**3.64**	**10.43**	**2.45**	**10.91**	**2.76**	**8.56**	**3.45**

Table 1 compares segmentation accuracy, boundary quality, and efficiency. S^3-Adapt obtains the highest Dice and mIoU on all four datasets. Relative to the strongest Dice competitor on each dataset, the gains are 0.18, 5.08, 1.23, and 0.29 points on QaTa-COV19, MosMedData+, Kvasir-SEG, and ISIC2016, respectively. It also ranks first in 15 of the 16 dataset–metric pairs. The sole exception is QaTa-COV19 HD95, where 12.80 is close to DIT's 12.71; meanwhile, ASSD improves from 3.76 to 3.64, indicating better average boundary agreement.

These gains use 1.97M trainable parameters, the second-lowest count after TextSAM-EUS and less than one-twentieth of those used by LViT, HiMix, or RecLMIS. Compared with TextSAM-EUS, S^3-Adapt adds 0.28M parameters, improves Dice and mIoU on every dataset, and reduces FLOPs from 1083.89G to 114.88G. Its FLOPs are not the lowest because both foundation encoders remain in the forward path; the efficiency advantage therefore concerns task-specific adaptation capacity rather than minimum computation.

4.3 Qualitative Analysis

Figure 5 compares five representative cases spanning chest X-ray, CT, endoscopy, and dermoscopy. In the two radiographs, several baselines activate broad or anatomically unrelated lung regions, whereas S^3-Adapt remains concentrated on the annotated infection, including the small unilateral

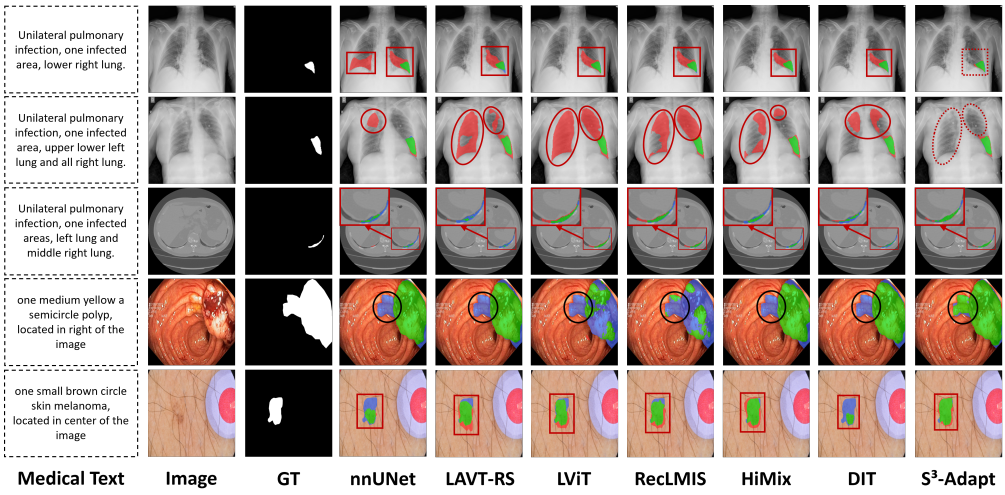

Figure 5: Qualitative comparison across four datasets. Green, red, and blue denote true-positive, false-positive, and false-negative regions, respectively.

target. The CT example contains a thin, curved lesion with weak contrast; our prediction preserves more of its extent while limiting surrounding false positives. For the polyp, competing masks either omit part of the target or absorb nearby mucosal folds, while S^3-Adapt better follows the visible boundary. The dermoscopic case similarly exposes over- and under-segmentation around an irregular lesion, for which our mask shows fewer interior omissions and less exterior spillover. Across these examples, the improvement reflects simultaneous suppression of false positives and recovery of missed target regions rather than uniform mask expansion or contraction. More cross-method visualizations are provided in Appendix C.

4.4 Ablation Studies

Table 2 evaluates RSG, HCST, and ECSC from a common similarity-and-prompt baseline with dense grounding, point/continuous-mask/text prompts, Decoder LoRA, OHA, and single-mask decoding. Unchanged modules share supervision and training budgets, and every ECSC variant refines a frozen reference. Additional sensitivity analyses for the learning rate and batch size are provided in Appendix D.

Table 2: Component ablation on MosMedData+ and Kvasir-SEG under shared training settings.

Configuration	RSG	HCST	ECSC	MosMedData+				Kvasir-SEG			
				Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
Baseline	–	–	–	71.96	58.31	14.73	3.94	88.65	81.62	27.84	8.48
+ RSG	✓	–	–	72.53	59.05	14.34	3.59	91.59	86.11	14.44	3.78
+ HCST	–	✓	–	75.46	62.38	12.43	3.15	92.33	<u>87.92</u>	<u>11.24</u>	3.55
+ ECSC	–	–	✓	73.47	60.09	13.75	3.32	91.79	86.41	14.38	4.54
+ RSG + HCST	✓	✓	–	<u>76.59</u>	<u>63.50</u>	10.76	<u>2.70</u>	<u>92.70</u>	87.84	11.93	3.59
+ HCST + ECSC	–	✓	✓	76.37	<u>63.50</u>	<u>10.68</u>	2.86	92.53	87.44	12.36	<u>3.37</u>
Full S^3-Adapt	✓	✓	✓	**77.74**	**65.03**	**10.43**	**2.45**	**93.08**	**88.47**	**10.91**	**2.76**

Component effects. The full model leads every metric, improving Dice by 5.78/4.43 and mIoU by 6.72/6.85 points over the baseline on MosMedData+ and Kvasir-SEG. RSG alone adds 0.57/2.94 Dice points, while HCST is the strongest single module with gains of 3.50/3.68 points, confirming that representation-level transfer contributes beyond prompts. Adding RSG to HCST provides another 1.13/0.37 points, showing that source preparation and hierarchical transfer are complementary.

Selective correction. Adding ECSC to RSG+HCST raises Dice by 1.15/0.38 and mIoU by 1.53/0.63 points, while reducing HD95 by 0.33/1.02 and ASSD by 0.25/0.83 pixels. Because the reference

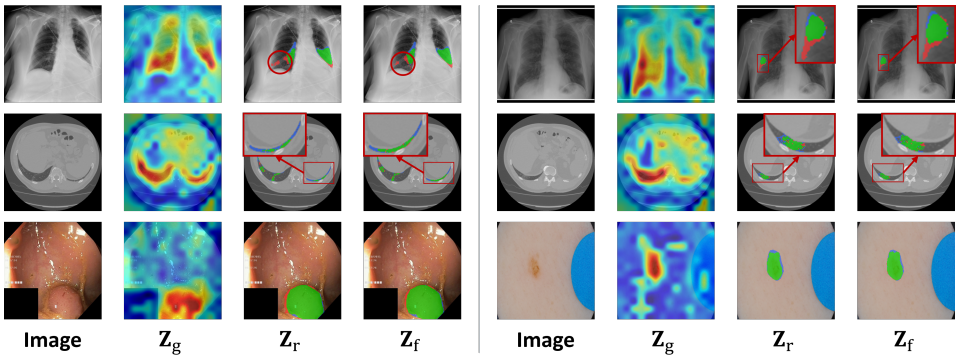

Figure 6: Selective-correction examples (left to right): image, grounding map $\mathbf{Z}_g$, reference prediction $\mathbf{Z}_r$, and final prediction $\mathbf{Z}_f$.

is frozen, these gains isolate correction from reference retraining. Figure 6 further shows that $\mathbf{Z}_g$ localizes the target coarsely and $\mathbf{Z}_r$ captures its main shape; $\mathbf{Z}_f$ then removes localized false positives or restores missed boundaries while preserving confident interiors. This behavior matches the bounded residual in Eq. 6 rather than global mask replacement.

Table 3: Design ablations for RSG/HCST placement and grounding supervision. Block indices are zero-based, arrows give initial/final grounding weights, and gray marks the default.

Category	Setting	MosMedData+				Kvasir-SEG			
		Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
RSG layers	[9]	75.85	63.01	11.18	3.23	92.66	88.21	11.28	3.37
	[7,9]	**77.74**	**65.03**	**10.43**	**2.45**	**93.08**	**88.47**	**10.91**	**2.76**
	[5,7,9]	76.29	63.44	10.69	2.90	92.22	87.37	13.81	3.98
	[3,5,7,9]	77.40	64.51	12.51	2.89	92.56	87.90	11.43	3.71
HCST layers	[11]	74.10	60.99	11.60	3.17	91.85	86.89	12.20	3.94
	[9,11]	76.20	63.18	11.62	3.01	92.26	87.36	12.51	3.97
	[2,9,11]	76.21	63.49	11.91	3.05	92.88	88.37	11.34	3.26
	[0,2,9,11]	**77.74**	**65.03**	**10.43**	**2.45**	**93.08**	**88.47**	**10.91**	**2.76**
	[1,4,7,10]	77.24	64.59	10.65	3.00	92.49	87.87	11.70	3.80
Grounding supervision	$0.00 \rightarrow 0.00$	77.11	64.21	10.84	2.63	92.27	87.40	12.84	3.96
	$0.10 \rightarrow 0.10$	77.42	64.53	**10.34**	2.52	93.04	88.38	**9.59**	**2.45**
	$0.20 \rightarrow 0.05$	**77.74**	**65.03**	10.43	**2.45**	**93.08**	**88.47**	10.91	2.76
	$0.40 \rightarrow 0.10$	77.41	64.60	11.72	3.01	92.31	87.57	11.95	3.84

Design choices. Table 3 shows that RSG blocks $[7, 9]$ and stage-end HCST blocks $[0, 2, 9, 11]$ lead all eight metrics. Relative to blocks $[9]$ and $[11]$, they improve Dice by 1.89/0.42 and 3.64/1.23 points, respectively. Adding earlier RSG blocks is not monotonic, while stage-end HCST also exceeds the equally sized $[1, 4, 7, 10]$ setting by 0.50/0.59 Dice points; thus, selective grounding and stage alignment both matter. The default $0.20 \rightarrow 0.05$ schedule improves Dice by 0.63/0.81 points over no grounding supervision and maximizes Dice and mIoU. Constant 0.10 gives better Kvasir-SEG boundary distances but slightly lower Dice, whereas $0.40 \rightarrow 0.10$ is consistently worse, indicating that moderate supervision provides the best trade-off.

5 Conclusion

We presented S^3-Adapt, a parameter-efficient framework that transfers grounded spatial semantics from a medical VLM to a segmentation foundation model. RSG prepares query-related source representations, HCST injects them into the target hierarchy, and ECSC locally corrects a fixed reference. With 1.97M task-adapted parameters, S^3-Adapt achieves the best Dice and mIoU on all four datasets and leads 15 of 16 dataset–metric comparisons. Ablations show that hierarchical transfer contributes most, RSG complements HCST, and ECSC improves overlap and boundaries, validating grounded representation transfer beyond decoder-only prompting.

AI USE STATEMENT

Generative AI tools were used for language editing, LaTeX assistance, and consistency checks of the manuscript and reported tables. They were not used to generate experimental data or execute experiments. The authors independently verified all AI-assisted revisions against the source results and literature and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

The supplementary material provides source code, configurations, and execution instructions. Section 4.1 and Appendices A–D specify the architecture, training stages, prompt construction, dataset splits, metrics, hyperparameters, software environment, and ablation settings. All experiments use fixed splits, seed 42, deterministic execution options, and identical preprocessing and metric implementations. These materials support reproduction of the reported results in the specified environment.

REFERENCES

Phuoc-Nguyen Bui, Duc-Tai Le, and Hyunseung Choo. Visual-textual matching attention for lesion segmentation in chest images. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 702–711, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72114-4.

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 Leonid Karlinsky, Tomer Michaeli, and Ko Nishino (eds.), *Computer Vision – ECCV 2022 Workshops*, pp. 205–218, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-25066-8.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. doi: 10.1016/j.media.2024.103280. URL https://doi.org/10.1016/j.media.2024.103280.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16664–16678, 2022.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3359–3367, 2023. doi: 10.1109/ICCVW60793.2023.00361.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 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, 2018. doi: 10.1109/ISBI.2018.8363547.

Aysen Degerli, Serkan Kiranyaz, Muhammad E. H. Chowdhury, and Moncef Gabbouj. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2306–2310, 2022. doi: 10.1109/ICIP46576.2022.9897412.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931, 2021. doi: 10.1109/ICCV48922.2021.00391.

Xiaorui Huang, Gen Luo, Chaoyang Zhu, Bo Tong, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Deep instruction tuning for segment anything model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 905–914, 2024.

Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(4):1821–1835, 2025. doi: 10.1109/TMI.2024.3523333.

Soojin Hwang, Jaeyoon Sim, and Won Hwa Kim. Himix : Hierarchical visual-textual mixing network for lesion segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5332–5341, March 2026.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Shayan Jalilian and Abdul Bais. Clip-guided sam: Parameter-efficient semantic conditioning for promptable segmentation, 2026. URL https://arxiv.org/abs/2605.24807.

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 Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve (eds.), *MultiMedia Modeling*, pp. 451–462, Cham, 2020. Springer International Publishing. ISBN 978-3-030-37734-2.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *Medical Image Analysis*, 106:103749, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103749. URL https://www.sciencedirect.com/science/article/pii/S1361841525002968.

Taha Koleilat, Hojat Asgariandehkordi, Omid Nejati Manzari, Berardino Barile, Yiming Xiao, and Hassan Rivaz. Medclipseg: Probabilistic vision-language adaptation for data-efficient and generalizable medical image segmentation. *arXiv preprint arXiv:2602.20423*, 2026.

Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective sam combination for open-vocabulary semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26081–26090, 2025. doi: 10.1109/CVPR52734.2025.02429.

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, 2024. doi: 10.1109/TMI.2023.3291719.

Yungeng Liu, Xuanzi Fang, Haijin Zeng, Qi Dai, and Yongyong Chen. Decoupling language guidance from backbones for text-guided medical segmentation, 2026. URL https://arxiv.org/abs/2607.09481.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos, 2025. URL https://arxiv.org/abs/2504.03600.

Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzymyrskyy, Natalya V Ledikhova, Victor A Gombolevskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

Qingtao Pan, Zhengrong Li, Guang Yang, Qing Yang, and Bing Ji. EviVLM: When evidential learning meets vision language model for medical image segmentation. *IEEE Transactions on Medical Imaging*, 45(4):1369–1382, 2026. doi: 10.1109/TMI.2025.3622492.

Kanchan Poudel, Manish Dhakal, Prasiddha Bhandari, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Exploring transfer learning in medical image segmentation using vision-language models. In Ninon Burgos, Caroline Petitjean, Maria Vakalopoulou, Stergios Christodoulidis, Pierrick Coupe, Hervé Delingette, Carole Lartizien, and Diana Mateus (eds.), *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 1142–1165. PMLR, 03–05 Jul 2024. URL https://proceedings.mlr.press/v250/poudel24a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Pascal Spiegler, Taha Koleilat, Arash Harirpoush, Corey S. Miller, Hassan Rivaz, Marta Kersten-Oertel, and Yiming Xiao. TextSAM-EUS: Text prompt learning for SAM to accurately segment pancreatic tumor in endoscopic ultrasound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 959–968, 2025.

Shiting Xiao, Rishabh Kabra, Yuhang Li, Donghyun Lee, Joao Carreira, and Priyadarshini Panda. OpenWorldSAM: Extending SAM2 for universal image segmentation with language prompts. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1193. URL https://papers.nips.cc/paper_files/paper/2025/hash/32fa3ddb3a6350cd1ab68e25f1c59982-Abstract-Conference.html.

Zhao Yang, Jiaqi Wang, Xubing Ye, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Language-aware vision transformer for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5238–5255, 2025. doi: 10.1109/TPAMI.2024.3468640.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025a. doi: 10.1056/AIoa2400640.

Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model, 2025b. URL https://arxiv.org/abs/2406.20076.

Longxuan Zhao, Tao Wang, Xinlin Zhang, Yuanbin Chen, Ence Yang, and Tong Tong. Mtgt: Multiscale text feature-guided transformer in medical image segmentation. *Image and Vision Computing*, 165:105846, 2026. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2025.105846. URL https://www.sciencedirect.com/science/article/pii/S0262885625004342.

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 Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 724–733, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43901-8.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 696–712, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19815-1.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi (eds.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00889-5.

A METHOD AND OPTIMIZATION DETAILS

A.1 ENCODER ADAPTATION AND GRADIENT BOUNDARIES

RSG is inserted into zero-indexed blocks 7 and 9 of the BiomedCLIP visual encoder, whose image-patch grid is 14×14; text associations exclude padding tokens. RSG and HCST use independent parameters at each block or level and share a bottleneck width of $d = 64$. The RSG query transformation comprises a depthwise convolution, GELU, and a pointwise convolution, while both HCST gate and candidate functions are $3d \to d \to d$ MLPs. The temperature is positively parameterized, update scales are bounded learnable signed quantities, and output projections use small-magnitude initialization. Table A.1 lists the HCST configuration.

Table A.1: HCST configuration for a 512×512 segmentation input.

Level	Native blocks	Insertion block	Spatial grid	Channels
1	0	0	128×128	96
2	1–2	2	64×64	192
3	3–9	9	32×32	384
4	10–11	11	16×16	768

In the first stage, original parameters remain frozen, but gradients are preserved along adaptation pathways. The reference loss backpropagates through HCST to $\mathbf{S}$ and RSG, the grounding loss constrains source representations through the frozen visual projection, and continuous-mask prompts retain input gradients; discrete point extraction is non-differentiable. The second stage fixes all reference parameters and training-mode states and updates only ECSC. Spatial resampling uses bilinear interpolation with `align_corners=False`.

A.2 PROMPT INTERFACE AND REFERENCE DECODING

Points and continuous masks. The maximum and minimum responses on the native cosine-similarity grid define one positive and one negative point, respectively. They are mapped to SFM input coordinates through image-patch centers, with ties resolved by fixed index order. The continuous mask is obtained by interpolating $\mathbf{Z}_g^{14}$ to the native prompt-encoder resolution without binarization. Training and inference use identical rules, and no ground-truth mask generates an input prompt.

Text tokens. Affine-free layer normalization and a bias-free projection transform the global text vector into $\mathbf{p}_b^t = W_t \,\mathrm{LN}_0(\mathbf{t}_b)$, where $W_t \in \mathbb{R}^{256\times512}$. The resulting $\mathbf{P}_t \in \mathbb{R}^{B\times1\times256}$ is appended to the native sparse point embeddings to form $\mathbf{P}_\mathrm{s}$; the continuous mask is processed into $\mathbf{P}_\mathrm{d}$. The native prompt encoder remains frozen.

Native decoding and parameter adaptation. The decoder receives $\mathbf{F}_{32} \in \mathbb{R}^{B\times256\times32\times32}$, $\mathbf{F}_{128} \in \mathbb{R}^{B\times32\times128\times128}$, and $\mathbf{F}_{64} \in \mathbb{R}^{B\times64\times64\times64}$, without a feature-recovery branch. Rank-8 LoRA applies the increment $r^{-1}B_L A_L \mathbf{x}$, with B_L zero-initialized, to the query, key, value, and output projections in decoder self-attention, bidirectional cross-attention, final token-to-image attention, and to MLP linear mappings (Hu et al., 2022). FSM fixes single-mask output and disables stability fallback. OHA updates the native mask tokens, output upsampling, hypernetwork MLPs, and two high-resolution feature projections; all other native decoder parameters remain frozen.

A.3 LOSSES AND TRAINING STATES

Let z_{bi} and y_{bi} denote the logit and label of sample b at pixel i, let $p_{bi} = \sigma(z_{bi})$, and let ℓ_{bi} denote the pixel-wise BCE computed from logits. Foreground and background terms are separately normalized over the entire batch, with weights $w_f = w_b = 0.5$:

$$\mathcal{L}_\mathrm{BCE} = w_f \frac{\sum_{b,i} y_{bi}\ell_{bi}}{\sum_{b,i} y_{bi} + \epsilon_B} + w_b \frac{\sum_{b,i}(1-y_{bi})\ell_{bi}}{\sum_{b,i}(1-y_{bi}) + \epsilon_B}. \tag{A.1}$$

Let the pixel weight be $\omega_{bi} = w_b + (w_f - w_b)y_{bi}$, and define weighted predictions and labels as $\widetilde{p}_{bi} = \omega_{bi}p_{bi}$ and $\widetilde{y}_{bi} = \omega_{bi}y_{bi}$. The squared soft Dice loss is computed per image and averaged over

the batch:

$$\mathcal{L}_{\mathrm{Dice}} = \frac{1}{B}\sum_{b=1}^{B}\left(1 - \frac{2\sum_i \widetilde{p}_{bi}\widetilde{y}_{bi} + \epsilon_s}{\sum_i(\widetilde{p}_{bi}^2 + \widetilde{y}_{bi}^2) + \epsilon_s + \epsilon_D}\right). \tag{A.2}$$

Here, ϵ_B and ϵ_D stabilize the denominators, and ϵ_s is a smoothing term. The grounding weight is initially constant and then follows a cosine decay. The second stage optimizes only the final segmentation objective. Table A.2 summarizes the parameter states.

Table A.2: Parameter states in the two training stages.

Parameter group	Reference stage	Correction stage
Both original encoder backbones, the native prompt encoder, and unselected decoder parameters	Frozen	Frozen
RSG, HCST, text projection, and Decoder LoRA	Trained	Fixed
OHA-selected native output parameters	Trained	Fixed
ECSC	Inactive	Trained

B Data and Evaluation Protocol

B.1 Data Organization and Text Sources

Image identifiers establish one-to-one image–mask–text triplets. Binary masks are stored as 0/255 and converted to 0/1 at loading; resizing, normalization, and augmentation occur in the input pipelines. Table B.1 reports two-dimensional images or slices rather than independent patients. Dataset sources appear in Section 4.1.

Table B.1: Dataset splits, imaging targets, and text sources used in the experiments.

Dataset	Modality / target	Train	Val	Test	Total	Text source
QaTa-COV19	Chest X-ray / infection	5,716	1,429	2,113	9,258	LViT
MosMedData+	Chest CT / lung infection	2,183	273	273	2,729	LViT
Kvasir-SEG	Endoscopy / polyp	800	100	100	1,000	MedVLSM P6
ISIC2016	Dermoscopy / skin lesion	810	90	379	1,279	MedVLSM P6
Total	–	9,509	1,892	2,865	14,266	–

QaTa-COV19 and MosMedData+ use LViT's fixed splits and image-specific infection texts. MosMedData+ is a processed two-dimensional version of MosMedData, so its slice count differs from the number of original CT examinations (Morozov et al., 2020; Li et al., 2024). Kvasir-SEG and ISIC2016 use MedVLSM P6 structured descriptions; when several candidates are available, the first is selected consistently (Poudel et al., 2024). These descriptions may specify category, count, size, shape, and coarse location.

B.2 Evaluation Metrics

For test image n, let P_n and Y_n denote the predicted and ground-truth foreground sets on the 224×224 grid. The per-image overlap scores are

$$\mathrm{Dice}_n = \frac{2|P_n \cap Y_n|}{|P_n| + |Y_n|}, \qquad \mathrm{IoU}_n = \frac{|P_n \cap Y_n|}{|P_n \cup Y_n|}. \tag{B.1}$$

Dice and mIoU are the arithmetic means over the N_{test} test images:

$$\mathrm{Dice} = \frac{1}{N_{\mathrm{test}}}\sum_{n=1}^{N_{\mathrm{test}}}\mathrm{Dice}_n, \qquad \mathrm{mIoU} = \frac{1}{N_{\mathrm{test}}}\sum_{n=1}^{N_{\mathrm{test}}}\mathrm{IoU}_n. \tag{B.2}$$

Let ∂P_n and ∂Y_n be the predicted and ground-truth boundary point sets, and define the directed nearest-boundary distance as

$$d(\mathbf{x}, \mathcal{B}) = \min_{\mathbf{b} \in \mathcal{B}} \|\mathbf{x} - \mathbf{b}\|_2. \tag{B.3}$$

The bidirectional boundary-distance multiset is

$$\mathcal{D}_n = \{d(\mathbf{p}, \partial Y_n) : \mathbf{p} \in \partial P_n\} \uplus \{d(\mathbf{y}, \partial P_n) : \mathbf{y} \in \partial Y_n\}, \tag{B.4}$$

where $\uplus$ denotes multiset union. The per-image 95th-percentile Hausdorff distance and average symmetric surface distance are

$$\mathrm{HD95}_n = Q_{0.95}(\mathcal{D}_n), \qquad \mathrm{ASSD}_n = \frac{\sum_{\mathbf{p} \in \partial P_n} d(\mathbf{p}, \partial Y_n) + \sum_{\mathbf{y} \in \partial Y_n} d(\mathbf{y}, \partial P_n)}{|\partial P_n| + |\partial Y_n|}, \tag{B.5}$$

where $Q_{0.95}$ is the 95th percentile. Reported HD95 and ASSD are the corresponding per-image means. All methods share the same boundary extraction, connectivity, binarization, and distance computation; distances are in pixels.

C Additional Qualitative Analysis

C.1 Extended Cross-Method Comparison

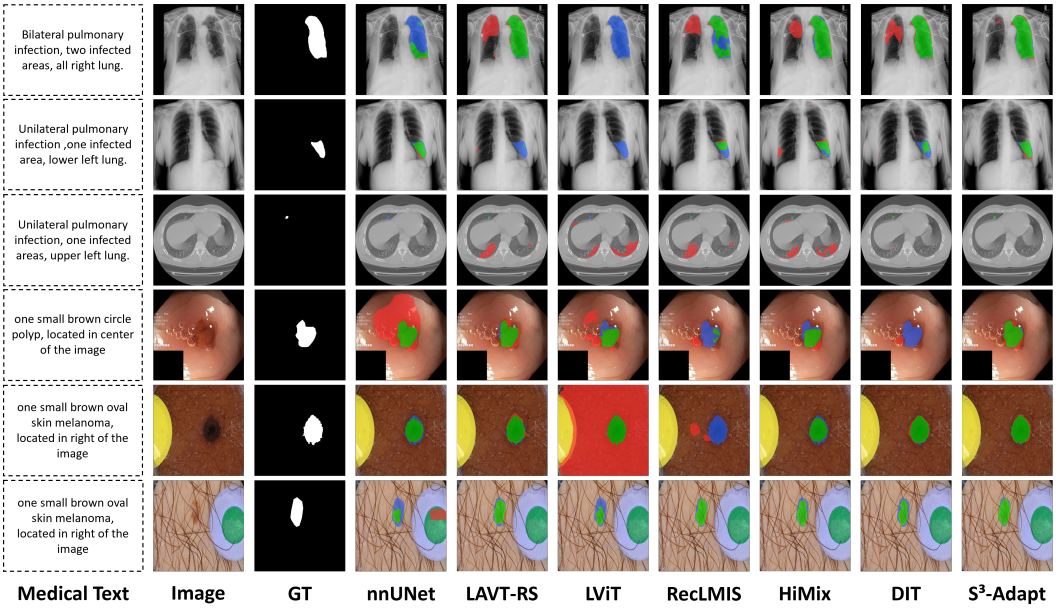

Figure 7: Additional qualitative comparisons across QaTa-COV19, MosMedData+, Kvasir-SEG, and ISIC2016.

Figure 7 adds two radiographs, one CT slice, one polyp, and two dermoscopic lesions. On radiographs, S^3-Adapt retains weak infection regions while suppressing broad false positives. The CT target occupies only a few pixels: several methods miss it or oversegment bilateral lung regions, whereas ours remains localized. The endoscopic result covers the polyp without absorbing nearby folds, and the dermoscopic results follow the lesion contours without large exterior responses. Differences are clearest for small, low-contrast, or irregular targets, where grounded localization and hierarchical transfer are most valuable.

C.2 Grounding-to-Correction Process

Figure 8 exposes the component interfaces. The grounding map provides coarse localization: its maximum and minimum yield deterministic positive and negative points, while the full-valued map

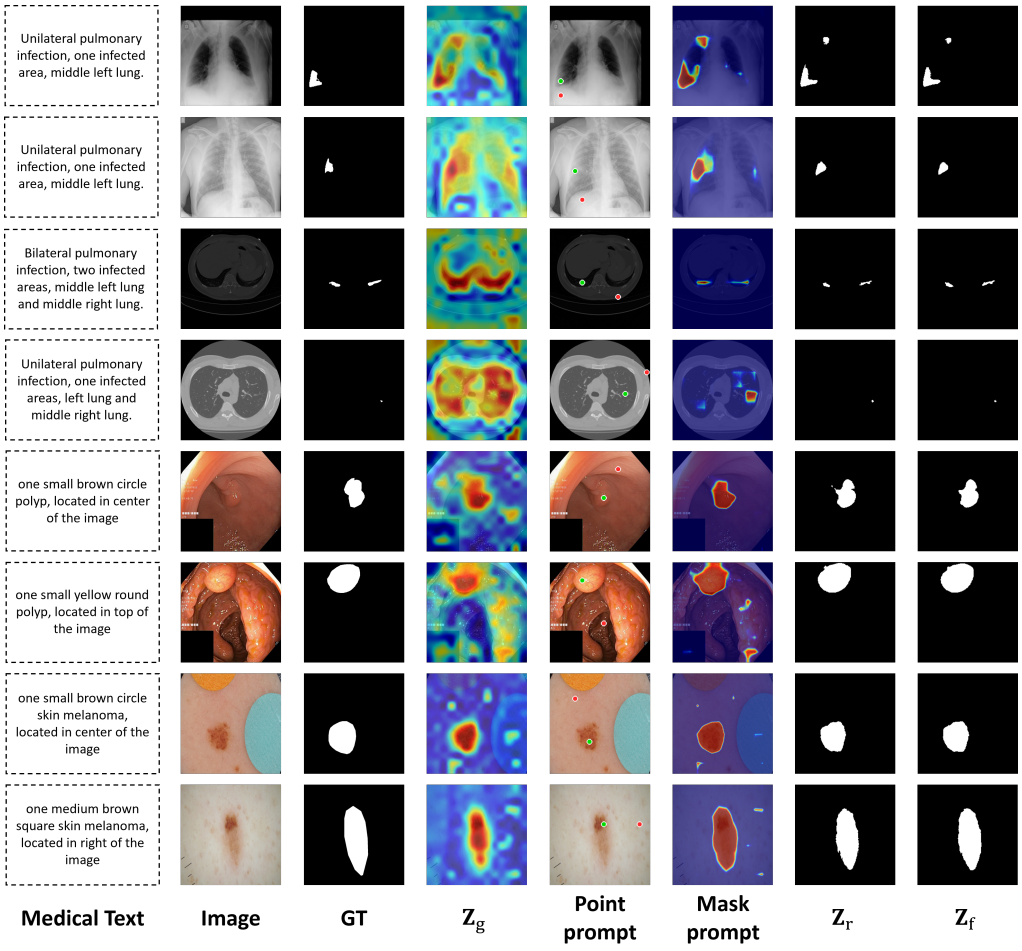

Figure 8: Prediction process from grounded prompts to reference and corrected masks.

supplies a dense mask prompt. The reference masks already capture the dominant target extent across modalities. Comparing $\mathbf{Z}_{\mathrm{r}}$ and $\mathbf{Z}_{\mathrm{f}}$ shows that ECSC makes localized changes to false-positive islands, missed boundaries, and small contour deviations instead of redrawing the mask globally. This behavior agrees with its uncertainty and grounding–reference disagreement inputs: RSG supplies spatial evidence, HCST and the prompts establish a strong reference, and ECSC corrects the residual errors.

D Hyperparameter Sensitivity

Table D.1: Learning-rate and batch-size sensitivity on MosMedData+, Kvasir-SEG, and ISIC2016. Gray indicates the default setting.

Group	Setting	MosMedData+				Kvasir-SEG				ISIC2016			
		Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
LR	10^{-4}	76.11	63.10	13.31	3.11	92.02	87.08	12.74	3.87	93.06	87.55	9.10	3.58
	3×10^{-4}	**77.74**	**65.03**	10.43	**2.45**	**93.08**	**88.47**	10.91	**2.76**	**93.14**	**87.71**	**8.56**	**3.45**
	5×10^{-4}	76.67	63.79	11.09	2.98	92.66	87.97	**10.51**	2.81	92.76	87.10	9.93	3.89
	10^{-3}	77.24	64.52	**10.06**	2.88	92.09	87.14	14.30	4.33	92.85	87.27	9.13	3.76
Batch	4	77.47	64.64	11.43	2.74	92.54	87.70	12.77	3.40	93.10	**87.73**	8.72	3.54
	8	**77.74**	**65.03**	**10.43**	**2.45**	**93.08**	**88.47**	**10.91**	**2.76**	**93.14**	87.71	**8.56**	**3.45**
	16	76.74	63.99	10.67	2.84	92.20	87.43	11.69	3.77	92.55	86.86	9.66	3.86
	32	76.30	63.15	12.11	3.02	92.06	87.08	12.65	3.95	92.59	86.92	9.98	3.92

Learning rate. Table D.1 shows that the default 3×10^{-4} gives the highest Dice and mIoU on all three datasets (77.74/65.03, 93.08/88.47, and 93.14/87.71) and the lowest ASSD. Boundary optima differ slightly: 10^{-3} gives the lowest MosMedData+ HD95 (10.06), and 5×10^{-4} gives the lowest Kvasir-SEG HD95 (10.51). The default nevertheless offers the most consistent overlap–boundary trade-off across datasets.

Batch size. A batch size of 8 gives the highest Dice and lowest HD95 and ASSD on all three datasets. The only exception among overlap metrics is ISIC2016 mIoU, where batch size 4 is higher by 0.02 (87.73 versus 87.71). Sizes 16 and 32 provide no improvement, supporting 8 as the most consistent setting.