LAYER-WISE SPATIAL EVIDENCE FOR MEDICAL VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical visual grounding requires identifying the intended lesion or anatomy and delineating its precise extent under limited pixel coverage, ambiguous boundaries, and cross-modal domain shift. Existing segmentation-token methods typically select a Transformer layer based on hidden states before constructing its spatial map. However, an isolated hidden state lacks the spatial context to reveal where its eventual spatial map will respond, resulting in suboptimal layer composition and background contamination. We introduce LaSE, the first medical grounding framework to adopt a *map-first* principle: it evaluates spatial evidence prior to layer fusion. LaSE comprises two novel components: (1) Layer-wise Token Matching (LTM), which constructs a candidate spatial map at every Transformer layer by matching the generated segmentation state with image tokens, exposing layer-wise spatial responses without pre-selecting a layer; and (2) Medical Evidence-Guided Fusion (MEGF), which pools features from the medical regions indicated by each spatial map in a shared frozen MedSAM2 representation and adaptively predicts case-dependent weights, thereby suppressing candidates that point to spurious anatomies. We further propose Lightweight Mask Refinement (LMR), which refines the decoded mask with a lightweight convolutional pyramid and requires only a single MedSAM2 decoder call. Extensive experiments on a four-source (Source-4), ten-target (OOD-10) medical segmentation benchmark demonstrate the effectiveness of LaSE: it achieves 90.20% Source-4 and 84.14% OOD-10 DSC, surpassing the state-of-the-art MedCLIPSeg by 1.09 and 5.12 points on Source-4 and OOD-10, respectively, while improving the pre-decoder map Soft Dice from 54.55% to 59.92%. Code is available at https://anonymous.4open.science/r/LaSE-22.

1 INTRODUCTION

Medical visual grounding aims to identify and delineate the region in a medical image referred to by a textual description, yielding a pixel-level segmentation mask. A successful response requires both semantic understanding to locate the intended lesion or anatomical structure, and precise spatial delineation to capture its extent. This is particularly challenging when targets are small, boundaries are ambiguous, and image appearance shifts drastically across modalities or institutions. Consequently, the central bottleneck lies not only in how a vision-language model comprehends the text, but in how this comprehension is translated into spatial evidence usable by a mask decoder.

Recent work has substantially strengthened different parts of this task. Disease-aware prompting improves weakly supervised localization without pixel annotations (Huy et al., 2025); MIMO connects visual referring inputs to phrase-grounded mask outputs (Chen et al., 2025); and MedSIGHT represents anatomical and pathological regions with generative region codes (Chang et al., 2026). MedReasoner further addresses implicit clinical queries by separating reasoning from expert segmentation (Yan et al., 2026). These approaches advance supervision, interaction, and grounded generation, but solve different problems from choosing *which* intermediate vision–language representations should form a spatial prompt. Text-guided medical segmentation, including MedCLIPSeg (Koleilat et al., 2026), also demonstrates the value of image–text alignment under domain shift; it does not resolve this layer-to-prompt choice in a segmentation-token architecture.

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.

Segmentation-token methods expose the choice more directly. LISA passes a generated segmentation state to a promptable decoder (Lai et al., 2024), whereas UGround explicitly matches that state to same-layer image tokens and uses the resulting spatial map to prompt SAM (Qian et al., 2026). UGround also learns to select a intermediate Transformer layer, but its policy acts on hidden representation *before* constructing the map of the selected layer. The selected map is spatial, yet spatial maps and the image regions they highlight cannot inform that prior decision. In medical images, a semantically plausible representation need not point to the intended tissue. This suggests a different ordering: expose each layer's spatial response first, then use the indicated medical-image evidence to decide how layers should contribute to the prompt.

To address this, we propose a *map-first* principle and instantiate it in LaSE. Instead of selecting layers based on hidden representation, LaSE first exposes the spatial maps of every layer. Specifically, the Layer-wise Token Matching (LTM) module constructs a candidate spatial map at each Transformer layer by matching the generated segmentation representation with image tokens. Subsequently, the Medical Evidence-Guided Fusion (MEGF) module pools features from the medical regions indicated by each spatial map in a shared, frozen MedSAM2 representation. This allows MEGF to adaptively predict case-dependent weights, ensuring that candidates pointing to spurious anatomies are suppressed. Weighted fusion then yields aligned spatial and semantic prompts, which drive a single MedSAM2 decoder call, followed by Lightweight Mask Refinement (LMR) to produce the final mask. Extensive experiments on the four-source (Source-4), ten-target (OOD-10) medical segmentation benchmark of Koleilat et al. (2026) show that LaSE reaches 90.20% Source-4 and 84.14% OOD-10 DSC, versus 89.11% and 79.02% in the *state-of-the-art* MedCLIPSeg comparison. Our contributions are threefold:

- We propose LaSE, the first medical visual grounding framework to evaluate spatial evidence prior to Transformer layer fusion. It introduces two core components: LTM for constructing candidate spatial maps at every layer to expose layer-wise spatial responses, and MEGF for adaptively weighting both the maps and their source token states, ensuring consistent multimodal prompt generation under a single-scale constraint.
- We further introduce Lightweight Mask Refinement (LMR), which refines the decoded mask with a lightweight convolutional pyramid and requires only a single MedSAM2 decoder call. This design reduces the computational overhead of multi-scale feature pyramids while maintaining high localization accuracy.
- Extensive experiments on the four-source, ten-target medical segmentation benchmark show the effectiveness of LaSE. Especially, LaSE achieves 90.20% Source-4 and 84.14% OOD-10 DSC, outperforming the state-of-the-art MedCLIPSeg by 1.09 and 5.12 points on Source-4 and OOD-10, respectively, with particularly strong gains on unseen targets. Moreover, the pre-decoder map Soft Dice improves from 54.55% to 59.92%, validating that the map-first ordering yields intrinsically better spatial cues.

2 Related Work

Medical Visual Grounding. The field spans distinct input, supervision, and output settings. Huy et al. (2025) improve weakly supervised disease localization in chest radiographs by emphasizing disease-relevant features without dense pixel labels. MIMO permits visual referring input and links generated medical phrases to output masks (Chen et al., 2025); MedSIGHT learns region-centric representations and generates codes that reconstruct grounded masks (Chang et al., 2026). MedReasoner couples reinforcement-learned reasoning about implicit clinical queries with a frozen segmentation expert (Yan et al., 2026), whereas MedLoc-R1 improves reinforcement-learning-based localization through adaptive reward scheduling (Yang et al., 2026). MedSG-Bench broadens the evaluation setting to grounding across medical image sequences (Yue et al., 2025). These settings are complementary to our single-image, pixel-level task: we study the interface between a vision–language representation and its spatial prompt, rather than sequence reasoning or reward design.

Text-conditioned Medical Segmentation. General-domain CLIPSeg and CAT-Seg establish text-conditioned mask prediction and dense image–text matching (Lüddecke & Ecker, 2022; Cho et al., 2024). In medical imaging, LViT integrates language into segmentation (Li et al., 2024), while CausalCLIPSeg addresses spurious image–text correlations in referring medical segmentation (Chen

et al., 2024). MedCLIPSeg uses probabilistic cross-modal adaptation and reports source-to-unseen-target transfer across ultrasound, MRI, dermoscopy, and endoscopy (Koleilat et al., 2026). These methods motivate robust medical grounding but do not evaluate spatial maps from *different multimodal Transformer depths* before composing a decoder prompt. Promptable segmenters such as SAM, MedSAM, and MedSAM2 provide the complementary image-to-mask interface (Kirillov et al., 2023; Ma et al., 2024; 2025).

Segmentation Tokens and Spatial Prompting. LISA projects a generated segmentation-token state into a SAM prompt (Lai et al., 2024); CoPRS obtains an explicit positional prior from a generated concentration token (Lu et al., 2026). Closest to our layer-wise spatial interface, UGround matches a [SEG] state to image tokens to form a similarity-map prompt and uses a learned policy to choose a Transformer layer (Qian et al., 2026). Its contribution establishes the utility of explicit map prompting, but the choice of layer precedes construction of that layer's map. In contrast, LASE exposes the full candidate bank before using regional MedSAM2 features to weight maps and their source states together. The distinction is the *evidence available when combining layers*, not the invention of spatial maps or promptable decoding.

3 METHODOLOGY

3.1 OVERVIEW OF LASE

Given a medical image $X^I \in \mathbb{R}^{3\times512\times512}$ and a target description X^P, LaSE aims to predict a pixel-level mask X^G. The core insight of LaSE is to *evaluate spatial evidence prior to Transformer layer fusion*, as an isolated hidden state lacks the spatial context to reveal where its eventual spatial map will respond. As illustrated in Figure 1, the LaSE decouples this process into four components: (i) MGemma, which extracts multimodal states; (ii) Layer-wise Token Matching (LTM), which exposes a spatial map at every layer; (iii) Medical Evidence-Guided Fusion (MEGF), which adaptively weights candidates using localized features; and (iv) a MedSAM2 decoder equipped with Lightweight Mask Refinement (LMR).

Specifically, for a generated [SEG] token, MGemma first extracts layer-wise control and image-token states, denoted as T_ℓ^{SEG} and T_ℓ^I, respectively. LTM subsequently computes their spatial correspondence maps $\{M_\ell\}$. To ground these candidates, MEGF maps each M_ℓ to the corresponding regions in the shared MedSAM2 features F^{SAM}, yielding regional summaries $\{u_\ell\}$. These summaries are aggregated to predict layer weights $\boldsymbol{\beta}$, which jointly compose a dense map M_F and a sparse control state T_F^{SEG}. Crucially, M_F and T_F^{SEG} serve as *prompts* rather than predicted masks: the MedSAM2 decoder and LMR are responsible for yielding the final segmentation X^G. Notably, a generated [REJ] token follows the identical decoding path but is overridden by an empty-mask post-processing.

3.2 MGEMMA:MEDGEMMA REPRESENTATIONS

MGemma uses the instruction-tuned MedGemma 1.5-4B backbone (Sellergren et al., 2026). Its $L = 34$ multimodal Transformer layers contextualize the image tokens with the target description. The user supplies X^P; MedGemma generates a response [SEG] for a present target or [REJ] for an empty target. We replay (X^I, X^P) to obtain, at each layer $\ell \in \{1, \ldots, L\}$, the image states $T_\ell^I = [t_{\ell,1}^I, \ldots, t_{\ell,N}^I]^\top \in \mathbb{R}^{N\times d}$ and the generated control-token state $T_\ell^{\mathrm{SEG}} \in \mathbb{R}^d$, where $N = 256$ and $d = 2560$. Textual context enters through these multimodal states; the subsequent matching uses the generated control state and image states, not separate text–image embeddings. For [REJ], the same readout takes the rejection-token state in the position denoted by T_ℓ^{SEG}, before the final empty-mask override.

3.3 LTM:LAYER-WISE TOKEN MATCHING

The layer is not selected before localization. Instead, LTM compares T_ℓ^{SEG} with every image token *at the same layer*:

$$S_{\ell,p} = (T_\ell^{\mathrm{SEG}})^\top t_{\ell,p}^I, \qquad p = 1, \ldots, N. \tag{1}$$

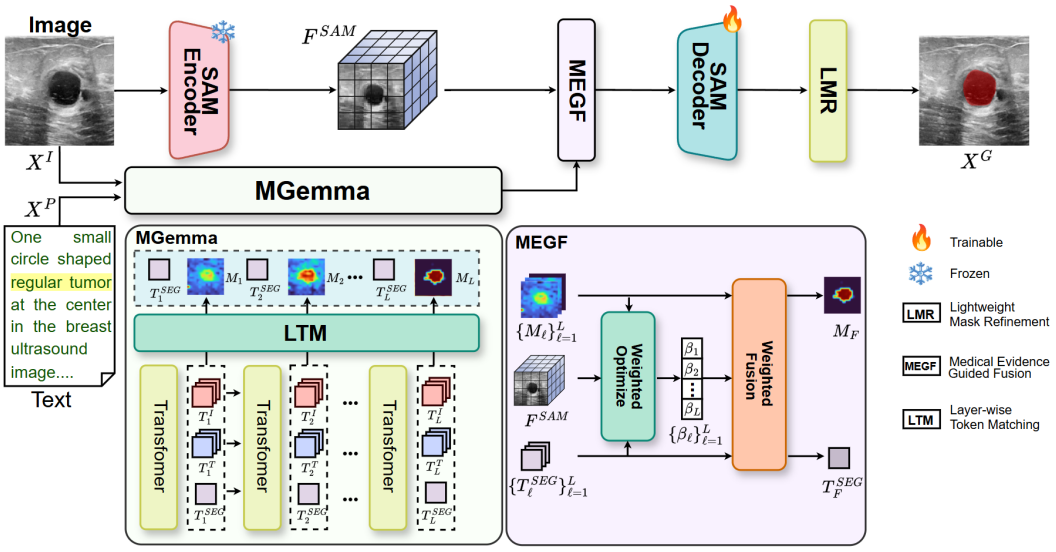


Figure 1: **Overview of LaSE for medical visual grounding.** LaSE takes an image X^I and a text prompt X^P to predict the segmentation mask X^G. A frozen MedSAM2 encoder extracts shared visual features F^{SAM}. Concurrently, the MGemma branch with Layer-wise Token Matching (LTM) converts layer-wise token states into candidate spatial maps $\{M_\ell\}$. To bridge coarse textual grounding with fine-grained visual features, the Medical Evidence-Guided Fusion (MEGF) module adaptively predicts layer weights $\{\beta_\ell\}$ by jointly aggregating $\{M_\ell\}$, [SEG] states $\{T_\ell^{\text{SEG}}\}$, and regional evidence from F^{SAM}. A weighted fusion synthesizes a dense map M_F and a sparse prompt T_F^{SEG}, which prompt a single trainable MedSAM2 decoder. Finally, the Lightweight Mask Refinement (LMR) module refines the decoder output into X^G. Flame/snowflake icons denote trainable/frozen modules, respectively.

Reshaping the $N = 256$ scores into a 16×16 grid and normalizing within each layer gives

$$M_\ell = \text{MinMax}_p\big(\text{reshape}(S_{\ell,1:N})\big) \in [0,1]^{1\times 16\times 16}. \tag{2}$$

M_ℓ measures *relative* image–token compatibility, rather than foreground probability. Unlike a layer-first strategy, this constructs all L spatial candidates before weighting them. The normalization is defined for a nonzero score range; non-finite or degenerate scores are treated as errors rather than silently replaced (Appendix A.4).

3.4 MEGF:Medical Evidence-Guided Fusion

An activated map proposes *where* to look, but its matching score alone does not establish whether that region contains the target. MEGF tests each proposal against a shared medical image representation. A frozen MedSAM2 image encoder with a SAM2.1 Hiera-tiny architecture (Ma et al., 2025) computes $(F^{\text{SAM}}, F^{\text{hi}}) = \xi_{\text{SAM2}}(X^I)$ once per image. Here $F^{\text{SAM}} \in \mathbb{R}^{256\times 32\times 32}$ evaluates candidate regions; multi-scale high-resolution features F^{hi} are retained for decoding. Resizing M_ℓ to the 32×32 grid yields $\widetilde{M}_\ell$. Its response-weighted regional feature is

$$a_{\ell,p} = \frac{\widetilde{M}_{\ell,p}}{\max\Big(\sum_{p'} \widetilde{M}_{\ell,p'}, \varepsilon\Big)}, \qquad u_\ell = \sum_p a_{\ell,p} F^{\text{SAM}}_{:,p}. \tag{3}$$

The sum ranges over feature-grid positions, $u_\ell \in \mathbb{R}^{256}$, and $\varepsilon = 10^{-8}$ guards against a zero-sum map. Different layers thus interrogate different candidate locations without re-encoding the image.

MEGF compares each candidate's semantic state with the medical features retrieved *by that candidate*. Separate projections of these signals feed a scorer shared across layers:

$$q_\ell = f_{\text{MEGF}}\big([\phi_{\text{SEG}}(T_\ell^{\text{SEG}}) \| \phi_{\text{med}}(u_\ell)]\big), \qquad \boldsymbol{\beta} = \text{entmax}_{1.5}\big((q_1,\ldots,q_L)/\tau\big). \tag{4}$$

Both ϕ_{SEG} and ϕ_{med} produce 128-dimensional features, $\|$ denotes concatenation, and f_{MEGF} is a two-hidden-layer MLP. A learned temperature $\tau \in [0.25, 8]$ controls the $\mathrm{entmax}_{1.5}$ transformation (Peters et al., 2019). It yields image–description-conditioned, potentially sparse weights $\beta_\ell \geq 0$ with $\sum_{\ell=1}^{L} \beta_\ell = 1$. Crucially, each q_ℓ depends on the spatial support of M_ℓ through u_ℓ: weighting can inspect a candidate's map rather than merely its layer index or hidden state.

Weighted Fusion. We use the *same* weights for spatial candidates and their source control states:

$$M_F = \sum_{\ell=1}^{L} \beta_\ell M_\ell, \qquad T_F^{\mathrm{SEG}} = \sum_{\ell=1}^{L} \beta_\ell T_\ell^{\mathrm{SEG}}. \tag{5}$$

Consequently, the dense cue $M_F \in [0,1]^{1\times 16\times 16}$ and sparse source $T_F^{\mathrm{SEG}} \in \mathbb{R}^d$ describe the same mixture of layer-wise evidence. No layer is committed to before its spatial response and regional support have been examined.

3.5 MedSAM2 Prompting and Mask Decoding

The two fused outputs enter complementary MedSAM2 prompt channels. The mask-prompt encoder consumes a resized M_F as a dense spatial cue, while a learned projection converts T_F^{SEG} into a sparse semantic token:

$$D_F = \mathcal{P}_{\mathrm{dense}}^{\mathrm{SAM2}}\big(\mathrm{Resize}(M_F)\big), \qquad E_F^{\mathrm{sp}} = [e_\varnothing \| \mathcal{P}_{\mathrm{sparse}}^{\mathrm{SEG}}(T_F^{\mathrm{SEG}})]. \tag{6}$$

$e_\varnothing$ is MedSAM2's empty-point embedding; the two prompt encoders do not threshold M_F or turn it into a predicted mask. Together with the shared image features and their positional encoding PE, these prompts produce low-resolution mask logits in one decoder call:

$$Z^{\mathrm{low}} = \mathcal{D}_{\mathrm{MedSAM2}}(F^{\mathrm{SAM}}, F^{\mathrm{hi}}, \mathrm{PE}, E_F^{\mathrm{sp}}, D_F). \tag{7}$$

We use the first mask-logit channel as Z^{low}; the final pixel mask is produced only after refinement.

3.6 Lightweight Mask Refinement

LTM and MEGF provide localization evidence; LMR addresses the distinct problem of delineating the decoded boundary. It interpolates the decoder logits and predicts a multi-scale correction from those same logits:

$$Z^{\mathrm{coarse}} = \mathrm{Resize}(Z^{\mathrm{low}}), \qquad \widehat{Z} = Z^{\mathrm{coarse}} + \Delta_{\mathrm{LMR}}(Z^{\mathrm{low}}), \qquad X^G = \sigma(\widehat{Z}). \tag{8}$$

Δ_{LMR} is a lightweight convolutional pyramid; $\widehat{Z}$ and X^G have 512×512 spatial resolution. LMR receives neither M_F nor $\boldsymbol{\beta}$ and does not call the decoder again. The coarse output, signed-distance field, boundary, scale, and interior consistency provide auxiliary supervision during training (Appendix A.6).

3.7 Training Objective

Each source-domain training pair (X^I, X^P) has a binary mask $Y \in \{0,1\}^{512\times 512}$. Its emptiness supplies the [SEG]/[REJ] target; for non-empty cases the mask also supervises the spatial prompt and segmentation. We optimize

$$\mathcal{L} = \mathcal{L}_{\mathrm{lang}} + \mathcal{L}_{\mathrm{map}} + \mathcal{L}_{\mathrm{mask}} + \mathcal{L}_{\mathrm{LMR}}^{\mathrm{aux}}. \tag{9}$$

$\mathcal{L}_{\mathrm{lang}}$ is response cross-entropy, including the control token. For non-empty cases, $\mathcal{L}_{\mathrm{map}}$ supervises M_F without specifying a correct layer, $\mathcal{L}_{\mathrm{mask}}$ supervises the refined logits $\widehat{Z}$, and $\mathcal{L}_{\mathrm{LMR}}^{\mathrm{aux}}$ supervises coarse and structural predictions. The last three terms are zero for empty targets, which contribute language loss only. Map supervision acts on a spatial *prompt*, not on a calibrated prediction; detailed targets and weights appear in Appendix A.4.

The MedGemma vision encoder and backbone remain frozen apart from query/value LoRA adapters (Hu et al., 2022); its tied input/output embedding matrix, including the added control-token rows, is trainable. The MedSAM2 image encoder is frozen, whereas MEGF, the prompt projections, the mask decoder, and LMR are optimized. At inference, only (X^I, X^P) is supplied: generation and replay select the control state, after which the map-first, single-decoder path above is applied. The rejection branch returns the registered empty mask; it does not use a target mask to make that decision (Appendix A.5).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Transfer protocol. We evaluate whether constructing layer-wise spatial evidence before fusion improves grounding under domain shift. Following the full-data MedCLIPSeg protocol (Koleilat et al., 2026), we train four separate source checkpoints: BUSI (breast ultrasound) (Al-Dhabyani et al., 2020), BTMRI (brain MRI) (Cheng, 2017), ISIC (dermoscopy) (Gutman et al., 2016), and Kvasir-SEG (endoscopy) (Jha et al., 2020). Each is tested on its source Test split and its registered unseen targets, yielding four source (Source-4) and ten OOD datasets (OOD-10): BUID, BUSBRA, BUSUC, UDIAT, BRISC, UWaterlooSkinCancer, BKAI, CVC300, ClinicDB, and ColonDB (Appendix A.1). Source validation alone determines the checkpoint; target images and masks are never used for model fitting, selection, or threshold tuning. Inputs X^P are the benchmark's released per-case descriptions with a fixed segmentation request, not newly collected clinical free text. Some descriptions contain coarse attributes derived from dataset annotations by the benchmark creators (Appendix A.2).

Performance Metrics. We report final-mask Dice similarity coefficient (DSC) to measure region overlap and normalized surface Dice (NSD) to assess boundary agreement within a two-pixel image-plane tolerance. Both are expressed as percentages. For our models, predicted probabilities are binarized strictly above 0.5 and evaluated in native image coordinates. Source-4 and OOD-10 are equal-dataset macro averages so that larger test sets do not dominate the summary. External comparisons use DSC because target-domain NSD was not reported for the published baselines. To examine the spatial cue before mask decoding, we additionally measure Soft Dice, which quantifies agreement between the continuous map M_F and the target, and foreground–background contrast, which measures its separation of target and background responses. These map diagnostics do not binarize M_F and are not final-mask scores (Appendix A.3).

Baselines. We compare LASE with the published results in the MedCLIPSeg benchmark (Koleilat et al., 2026). The comparison includes LViT (Li et al., 2024), CLIPSeg (Lüddecke & Ecker, 2022), DenseCLIP (Rao et al., 2022), ZegCLIP (Zhou et al., 2023), SAN (Xu et al., 2023), MaPLe (Khattak et al., 2023), VLSM-Adapter (Dhakal et al., 2024), CausalCLIPSeg (Chen et al., 2024), CAT-Seg (Cho et al., 2024), and MedCLIPSeg (Koleilat et al., 2026). MaPLe+Decoder is the decoder-augmented MaPLe variant reported in the same benchmark. We use the published DSC results for these methods and evaluate LASE on the same source–target matrix.

4.2 CROSS-DOMAIN GENERALIZATION

Table 1 reports the source-to-target DSC matrix. LASE reaches 90.20% Source-4 and 84.14% OOD-10 DSC, versus 89.11% and 79.02% for the published MedCLIPSeg row: a 5.12-point OOD-10 difference. LASE scores higher on all ten target datasets and is best on 12 of the 14 tests overall. The target gains span the four source families rather than coming from one domain: for example, BUID improves by 7.44%, UWaterlooSkinCancer by 8.39%, and BKAI by 9.90%. It does not lead on the BUSI and ISIC source tests, where MedCLIPSeg scores 0.94% and 0.25% points higher, respectively.

For the fixed LaSE checkpoints, 10,000 within-dataset case-bootstrap resamples yield 95% intervals of 88.77–91.47 (Source-4) and 82.58–85.56 (OOD-10) DSC. These intervals reflect case sampling for our model, not training-run variance or a paired significance test against the published baseline.

4.3 COMPONENT ABLATIONS

Table 2 compares LASE with two component variants. Source-4 and OOD-10 are equal-dataset means over four source test sets and ten unseen target test sets, respectively. The **w/o MEGF** variant replaces MEGF's map-conditioned fusion with *layer-first routing*: a categorical policy selects one layer from the layer-wise [SEG] states before constructing its spatial map. The policy is trained with REINFORCE, using a reward derived from the selected map's mask-supervised loss and an exponential-moving-average baseline. The selected map and token state provide the dense and sparse prompts, while LMR still refines the decoded mask. The **w/o LMR** variant constructs all spatial maps

Table 1: Domain generalization on four source and ten OOD datasets, measured by DSC (%).

	Breast ultrasound					Brain MRI		Polyp endoscopy					Dermatoscopy	
Method	BUSI	BUSBRA	BUSUC	BUID	UDIAT	BTMRI	BRISC	Kvasir	ColonDB	ClinicDB	CVC300	BKAI	ISIC	UWSC
LViT	75.32	59.41	67.95	53.51	65.60	81.41	71.86	85.29	60.01	75.27	70.22	67.17	91.21	58.87
CLIPSeg	80.95	63.66	75.03	68.43	56.67	86.33	77.61	81.98	59.93	71.49	72.74	66.46	90.55	80.19
DenseCLIP	71.85	53.34	70.97	63.53	54.93	70.30	34.12	79.32	56.38	68.08	64.71	61.63	89.29	53.39
ZegCLIP	72.08	61.08	73.57	71.75	52.41	76.65	66.31	78.46	53.46	69.75	60.73	65.60	81.45	38.60
SAN	77.99	64.37	74.15	58.13	61.98	85.27	71.60	83.16	61.82	74.46	80.36	69.31	91.39	82.51
MaPLe	66.37	50.08	71.52	70.77	57.81	75.40	45.19	76.12	48.09	59.64	63.80	56.94	88.31	69.12
MaPLe+Decoder	80.49	55.89	64.96	60.66	59.44	85.08	71.46	83.46	61.53	71.20	74.62	66.93	90.10	81.83
VLSM-Adapter	80.90	68.48	82.37	75.26	69.16	85.03	68.92	85.89	63.51	76.09	75.24	71.59	91.30	82.17
CausalCLIPSeg	76.11	55.87	69.12	64.49	48.90	81.71	53.96	78.77	41.65	57.54	45.77	52.56	89.47	48.73
CAT-Seg	81.83	70.94	81.48	73.37	70.30	84.86	76.28	86.43	68.49	70.35	78.12	74.35	91.27	82.02
MedCLIPSeg	**85.72**	75.06	84.37	78.99	74.64	88.03	80.92	90.15	71.90	80.80	80.82	79.15	**92.54**	83.53
LaSE	84.78	**75.97**	**85.65**	**86.43**	**78.51**	**90.71**	**82.81**	**93.00**	**79.33**	**84.35**	**87.35**	**89.05**	92.29	**91.92**

Note. UWSC denotes UWaterlooSkinCancer. Higher DSC is better; bold and underline mark the best and second-best result in each dataset column.

and retains MEGF's region-conditioned, shared-weight fusion, but removes final mask refinement. The full LASE system uses both MEGF and LMR.

Table 2: Component ablations results (%). Bold and underline mark the best and second-best value in each metric column, respectively.

	Source-4		OOD-10	
Variant	DSC	NSD	DSC	NSD
w/o MEGF	89.70	91.83	82.54	85.92
w/o LMR	89.87	92.01	82.81	86.29
LASE	**90.20**	**92.34**	**84.14**	**87.55**

LASE leads all four macro metrics. Its OOD-10 DSC is 84.14, compared with 82.54 for w/o MEGF. Removing LMR from the map-first system gives 82.81 OOD-10 DSC; restoring LMR raises it by 1.33 points. The two variants examine layer routing and final-mask refinement, respectively.

4.4 VISUALIZATION ANALYSIS

Pre-decoder Spatial Evidence: Layer-first vs. Map-first. To isolate whether the performance gain stems from improved spatial prompts or from the MedSAM2 decoder, we evaluate the pre-decoder fused map M_F against the map generated by the layer-first variant. This distinction is crucial because final-mask DSC alone can obscure the quality of the spatial cue before decoding. Table 3 reports the Soft Dice (foreground overlap) for this comparison. Quantitatively, the map-first strategy consistently outperforms the layer-first baseline. On OOD-10, it improves Soft Dice from 54.55% to 59.92% ($+5.37$ points). The Source-4 split exhibits the same directional gain, with Soft Dice increasing from 63.17% to 66.99%. These metrics characterize the continuous $[0, 1]$ spatial prompts before any thresholding or mask decoding (Appendix A.3), confirming that the map-first ordering yields intrinsically better spatial cues rather than relying on post-hoc mask refinement.

Figure 2 provides concrete visual evidence for this claim. Across all four OOD datasets, the layer-first maps exhibit fragmented responses, whereas the fused map M_F from LaSE responds more continuously and accurately covers the annotated targets. For instance, in the BUSBRA ultrasound case, the layer-first map presents low-response bands within the annotated target, while M_F maintains a continuous response. More notably, in the UWaterlooSkinCancer dermoscopy case, LaSE captures a significantly larger extent of the lesion, whereas the layer-first strategy misses its lower portion. The BRISC and BKAI columns confirm the same behavior on MRI and endoscopy, demonstrating cross-modal robustness. These visual patterns align with the quantitative gains in Table 3, indicating that evaluating spatial evidence before layer fusion effectively mitigates background contamination and preserves lesion completeness. Finally, the learned layer weights $\boldsymbol{\beta}$ are not fixed across source families, and their distributions shift from source to OOD targets (Appendix A.11), illustrating that the model dynamically composes candidates based on the input rather than relying on a statically optimal layer. This adaptive behavior provides direct insight into how LTM and MEGF cooperate:

Table 3: Soft Dice of LaSE on layer-first and map-first strategies (%). Bold and underline mark the higher and lower value per split.

	Layer-first		Map-first (Ours)	
Metric	Source-4	OOD-10	Source-4	OOD-10
Soft Dice	<u>63.17</u>	<u>54.55</u>	**66.99**	**59.92**

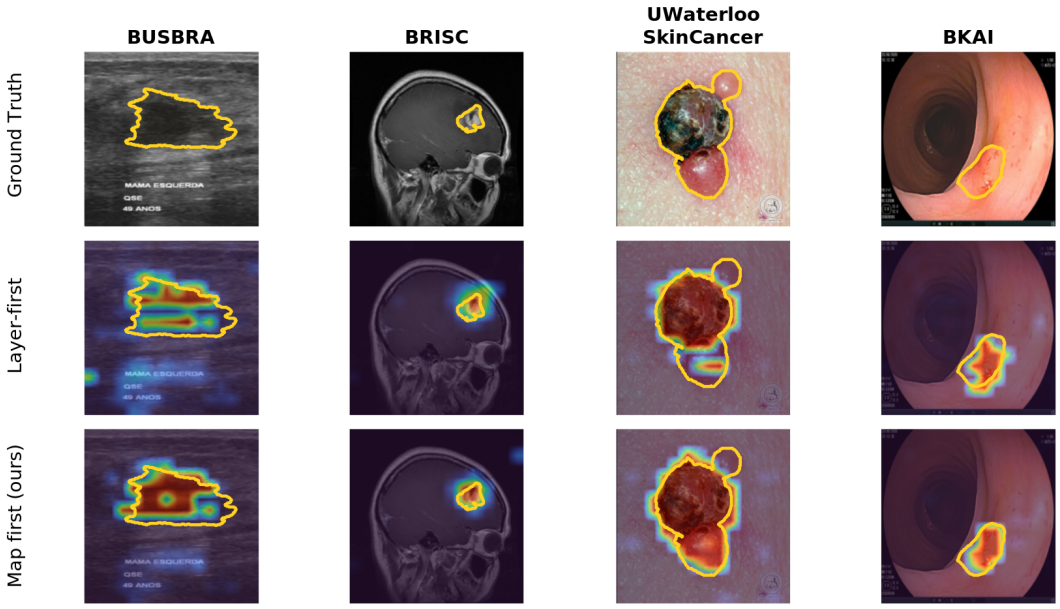


Figure 2: Case studies of LaSE with layer-first and map-first strategies on four OOD datasets.

LTM exposes all layer-wise candidates, while MEGF selectively amplifies those whose indicated regions align with medical evidence.

Mask visualization of LaSE with layer-first and map-first strategies on OOD-10 datasets. Figures 3 and 4 present qualitative mask comparisons of LaSE against MedCLIPSeg across the ten OOD-10 targets. By overlaying predictions on identical images, we disentangle the behavior of intermediate spatial prompts from the final masks produced by MedSAM2 and LMR. For instance, in the UWaterlooSkinCancer case, LaSE captures a larger lateral extent of the lesion than the MedCLIPSeg prediction. However, the BUID example reveals a residual disconnected off-target prediction, indicating that improved aggregate overlap does not fully eliminate localization errors, possibly due to ambiguous boundaries or persistent background clutter. Beyond final masks, Figure 5 illustrates the complete pipeline for four additional OOD cases, tracing the released description X^P through the fused prompt M_F to the final mask X^G. This end-to-end visualization highlights the interpretability of LaSE: the fused map provides a spatially coherent cue that MedSAM2 and LMR refine into the final segmentation. Additional pre-decoder examples are provided in Figures 12 and 13. Collectively, these qualitative results corroborate the quantitative findings in Tables 1 and 3, while also exposing remaining challenges such as occasional off-target predictions, which we leave for future work.

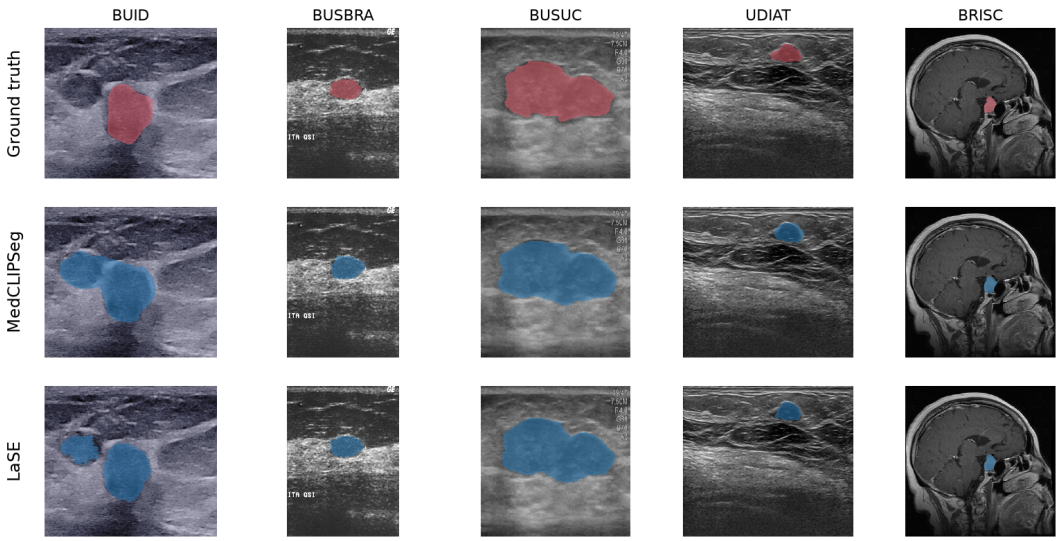


Figure 3: Visualization of generated masks of MedCLIPSeg and LaSE on BUID, BUSBRA, BUSUC, UDIAT, and BRISC.

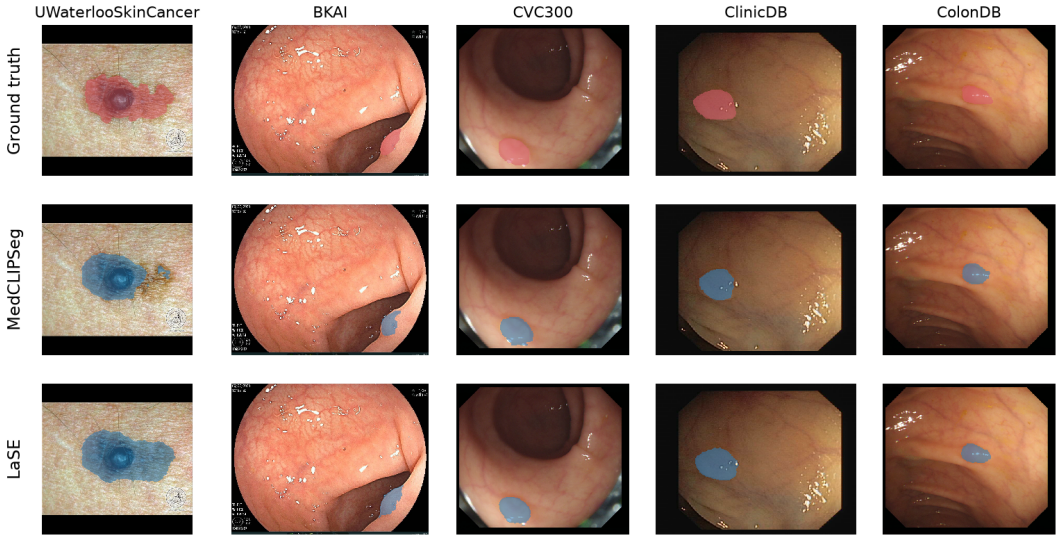


Figure 4: Visualization of generated masks of MedCLIPSeg and LaSE on UWaterlooSkinCancer, BKAI, CVC300, ClinicDB, and ColonDB.

5 Conclusion

In this work, we propose LaSE, an efficient medical visual grounding framework that addresses the lack of spatial evidence prior to Transformer layer fusion in segmentation-token paradigms. By constructing candidate spatial maps at every layer and adaptively weighting them using regional MedSAM2 evidence before composing a single decoder prompt, LaSE achieves superior cross-domain generalization with substantially reduced computational overhead. Limitation and Future Work: However, as analyzed in our experimental section, our model still has limitations: its performance is specific to the benchmark's released descriptions, some of which encode annotation-derived attributes, and the inference cost remains dominated by a large multimodal backbone. In future work, addressing these issues can improve the generalizability of LaSE to independently authored clinical requests and enable more efficient deployment. Meanwhile, the potential of applying the map-first principle to other multimodal segmentation architectures remains to be explored.

AI USE STATEMENT

We used large language models (LLMs) solely for polishing the writing, such as improving sentence fluency and fixing grammar issues. No content generation, idea creation, or experimental analysis was conducted by LLMs

ETHICS STATEMENT

This study uses previously released research datasets and involves no new patient recruitment or intervention. Dataset-specific access conditions and de-identification follow the source releases.

REPRODUCIBILITY STATEMENT

The model architecture is introduced in detail with equations and figures in the main text. Appendices A.1–A.11 document the dataset and prompt protocol, metric definitions, implementation details, inference contract, qualitative checks, and the β layer-weight distribution.Code is available at this repository :https://anonymous.4open.science/r/LaSE-22

REFERENCES

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. URL https://doi.org/10.1016/j.compbiomed.2022.106438.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. URL https://doi.org/10.1016/j.dib.2019.104863.

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, 2015. URL https://doi.org/10.1109/TBME.2014.2365518.

J. Bernal, J. Sánchez, and F. Vilariño. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012. URL https://doi.org/10.1016/j.patcog.2012.03.002.

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. URL https://doi.org/10.1016/j.compmedimag.2015.02.007.

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. URL https://doi.org/10.1016/j.bspc.2020.102027.

Aofei Chang, Le Huang, Alex James Boyd, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. MedSIGHT: Towards grounded visual comprehension in medical large vision-language models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/63100.

Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. MIMO: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24732–24741, 2025. URL https://doi.org/10.1109/CVPR52734.2025.02303.

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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15003 of *Lecture Notes in Computer Science*, pp. 77–87, Cham, 2024. Springer. URL https://doi.org/10.1007/978-3-031-72384-1_8.

Jun Cheng. brain tumor dataset. Figshare dataset, version 5, 2017. URL https://doi.org/10.6084/m9.figshare.1512427.v5.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4123, 2024. URL https://doi.org/10.1109/CVPR52733.2024.00394.

Manish Dhakal, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. VLSM-Adapter: Finetuning vision-language segmentation efficiently with lightweight blocks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15009 of *Lecture Notes in Computer Science*, pp. 712–722, Cham, 2024. Springer. URL https://doi.org/10.1007/978-3-031-72114-4_68.

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. URL https://doi.org/10.1038/s41597-026-06753-y.

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. URL https://doi.org/10.1109/TBME.2013.2244596.

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. URL https://doi.org/10.1002/mp.16812.

David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC), 2016. URL https://doi.org/10.48550/arXiv.1605.01397.

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.

Ta Duc Huy, Duy Anh Huynh, Yutong Xie, Yuankai Qi, Qi Chen, Phi Le Nguyen, Sen Kim Tran, Son Lam Phung, Anton van den Hengel, Zhibin Liao, Minh-Son To, Johan W. Verjans, and Vu Minh Hieu Phan. Seeing the trees for the forest: Rethinking weakly-supervised medical visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24445–24455, 2025. URL https://doi.org/10.1109/ICCV51701.2025.02266.

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 *MultiMedia Modeling*, pp. 451–462. Springer International Publishing, 2020. URL https://doi.org/10.1007/978-3-030-37734-2_37.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Khattak_MaPLe_Multi-Modal_Prompt_Learning_CVPR_2023_paper.html.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

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 (CVPR)*, pp. 1406–1417, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Koleilat_MedCLIPSeg_Probabilistic_Vision-Language_Adaptation_for_Data-Efficient_and_Generalizable_Medical_Image_CVPR_2026_paper.html.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lai_LISA_Reasoning_Segmentation_via_Large_Language_Model_CVPR_2024_paper.html.

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 *Advances in Visual Computing*, Lecture Notes in Computer Science, pp. 15–28. Springer, 2021. URL https://doi.org/10.1007/978-3-030-90436-4_2.

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. URL https://doi.org/10.1109/TMI.2023.3291719.

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Yan Feng, Bin Chen, Ke Chen, and Yaowei Wang. CoPRS: Learning positional prior from chain-of-thought for reasoning segmentation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 37940–37959, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/4048c740bec4c682b17c640f584d2762-Paper-Conference.pdf.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7086–7096, 2022. URL https://doi.org/10.1109/CVPR52688.2022.00695.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024. URL https://doi.org/10.1038/s41467-024-44824-z.

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.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1146. URL https://aclanthology.org/P19-1146/.

Rui Qian, Xin Yin, Chuanhang Deng, Zhiyuan Peng, Jian Xiong, Wei Zhai, and Dejing Dou. UGround: Towards unified visual grounding with unrolled transformers. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/65753.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18082–18091, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rao_DenseCLIP_Language-Guided_Dense_Prediction_With_Context-Aware_Prompting_CVPR_2022_paper.html.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. MedGemma technical report, 2026. URL https://arxiv.org/abs/2507.05201.

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*, 2017:1–9, 2017. URL https://doi.org/10.1155/2017/4037190.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2945–2954, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Xu_Side_Adapter_Network_for_Open-Vocabulary_Semantic_Segmentation_CVPR_2023_paper.html.

Zhonghao Yan, Muxi Diao, Yuxuan Yang, Ruoyan Jing, Jiayuan Xu, Kaizhou Zhang, Lele Yang, Yanxi Liu, Kongming Liang, and Zhanyu Ma. MedReasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11577–11585, 2026. URL https://doi.org/10.1609/aaai.v40i14.38141.

Guangjing Yang, Ziyuan Qin, Chaoran Zhang, Chenlin Du, Jinglin Wang, Wanran Sun, Zhenyu Zhang, Bing Ji, and Qicheng Lao. MedLoc-R1: Performance-aware curriculum reward scheduling for GRPO-based medical visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21036–21045, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_MedLoc-R1_Performance-Aware_Curriculum_Reward_Scheduling_for_GRPO-Based_Medical_Visual_Grounding_CVPR_2026_paper.html.

Moi Hoon Yap, Gerard Pons, Joan Martí, Sergi Ganau, Melcior Sentís, Reyer Zwiggelaar, Adrian K. Davison, and Robert Martí. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 22(4):1218–1226, 2018. URL https://doi.org/10.1109/JBHI.2017.2731873.

Jingkun Yue, Siqi Zhang, Zinan Jia, Huihuan Xu, Zongbo Han, Xiaohong Liu, and Guangyu Wang. MedSG-Bench: A benchmark for medical image sequences grounding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2502. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6c22b0280d0d7d9e0c3126b1e92df1a0-Abstract-Datasets_and_Benchmarks_Track.html.

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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Lecture Notes

in Computer Science, pp. 724–733. Springer, 2023. URL https://doi.org/10.1007/978-3-031-43901-8_69.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. ZegCLIP: Towards adapting CLIP for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11175–11185, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhou_ZegCLIP_Towards_Adapting_CLIP_for_Zero-Shot_Semantic_Segmentation_CVPR_2023_paper.html.

A Appendix

A.1 Dataset and Evaluation Protocol

We train one source-specific model on BUSI (Al-Dhabyani et al., 2020), BTMRI (Cheng, 2017), ISIC (Gutman et al., 2016), and Kvasir-SEG (Jha et al., 2020). Each checkpoint is selected on its source validation split and then evaluated on the registered source Test set and unseen targets. spatial maps are first restored to the canonical 512×512 evaluation grid and final predictions are then mapped back to each dataset's native coordinates before mask evaluation.

The registered OOD records are BUID (Abbasian Ardakani et al., 2023), BUSBRA (Gómez-Flores et al., 2024), BUSUC (Koleilat et al., 2026), UDIAT (Byra et al., 2020; Yap et al., 2018), BRISC (Fateh et al., 2026), UWaterlooSkinCancer (Amelard et al., 2015; Glaister et al., 2013), BKAI (Lan et al., 2021), CVC300 (Vázquez et al., 2017), ColonDB (Bernal et al., 2012), and ClinicDB (Bernal et al., 2015). Table 4 reports the image counts in the released benchmark splits and the source checkpoint used for each Test set. For both source and target cases, text inputs come from the `Description` column of the MedCLIPSeg benchmark's released prompt spreadsheets (Koleilat et al., 2026).

For a dataset-level metric m_d, the reported macros are

$$\text{Source-4} = \frac{1}{4} \sum_{d \in \mathcal{D}_{\text{source}}} m_d, \qquad \text{OOD-10} = \frac{1}{10} \sum_{d \in \mathcal{D}_{\text{OOD}}} m_d. \tag{10}$$

They are not weighted by dataset size. Public MedCLIPSeg target results do not include NSD, so external comparisons report DSC consistently.

Published comparison methods. Table 1 reproduces the published DSC matrix from MedCLIPSeg (Koleilat et al., 2026). Its comparison rows cover LViT, CLIPSeg, DenseCLIP, ZegCLIP, SAN, MaPLe, MaPLe+Decoder, VLSM-Adapter, CausalCLIPSeg, CAT-Seg, and MedCLIPSeg. We do not rerun these published baselines; only the LASE and component-control rows use local checkpoints.

Table 4: Released benchmark split sizes (image counts). Source Train/Val splits are used for fitting/checkpoint selection; OOD Train/Val splits are listed for dataset inventory only and are not used. Each Test set is evaluated by the source checkpoint in the final column.

Dataset	Role	Train	Val	Test	Total	Checkpoint
BUSI	Source	624	78	78	780	BUSI
BTMRI	Source	2738	1321	1005	5064	BTMRI
ISIC	Source	809	90	379	1278	ISIC
Kvasir-SEG	Source	800	100	100	1000	Kvasir
BUID	OOD	162	35	35	232	BUSI
BUSBRA	OOD	1311	282	282	1875	BUSI
BUSUC	OOD	567	122	122	811	BUSI
UDIAT	OOD	113	25	25	163	BUSI
BRISC	OOD	4000	1000	1000	6000	BTMRI
UWaterlooSkinCancer	OOD	132	0	41	173	ISIC
BKAI	OOD	799	100	100	999	Kvasir
CVC300	OOD	6	0	60	66	Kvasir
ClinicDB	OOD	490	61	61	612	Kvasir
ColonDB	OOD	20	0	360	380	Kvasir

A.2 Prompt provenance and evaluation scope

The benchmark's published prompt-generation procedure uses dataset labels and, when available, mask-derived coarse attributes such as location, component count, and shape (Koleilat et al., 2026). At inference, we append the fixed request "Please output segmentation mask." to each released description. The model receives this text and the image, without a target-mask tensor.

A.3 Mask and similarity-map evaluation

For LASE and the matched component variants, the final sigmoid probability is binarized with the strict threshold > 0.5 on the 512×512 evaluation grid, then the binary prediction is restored to the native image shape by nearest-neighbor interpolation. DSC and NSD are computed against the native-coordinate ground-truth mask. The separately reconstructed MedCLIPSeg masks shown in qualitative figures are resized from its hard 224×224 predictions; they do not replace the published scores in Table 1. When both prediction and ground truth are empty, DSC and NSD are 1; when only one is empty, both are 0. NSD uses the published MedCLIPSeg surface-distance implementation with unit voxel spacing and a two-pixel tolerance on the image plane, not patient-specific physical spacing. For either metric, m_d in the macro formula above is the mean over all registered cases of dataset d, including empty targets.

The similarity-map diagnostic uses a separate Native-512 protocol. Each 16×16 spatial map M_ℓ is independently min–max normalized to $[0, 1]$ during construction. The fused map $M_F = \sum_{\ell=1}^{L} \beta_\ell M_\ell$ is a convex combination of these spatial maps and is not min–max normalized again. For evaluation, each candidate or fused map is bilinearly resized to 512×512 with `align_corners=false`, without a sigmoid, clipping, or further normalization. The ground-truth mask is nearest-neighbor resized to the same grid and binarized at > 127. For map M and binary foreground Y, the reported Soft Dice and foreground– background contrast are

$$
\begin{aligned}
D_{\mathrm{soft}}(M, Y) &= \frac{2 \sum_p M_p Y_p + 10^{-8}}{\sum_p M_p + \sum_p Y_p + 10^{-8}}, \\
C_{\mathrm{FG-BG}}(M, Y) &= \frac{\sum_p M_p Y_p}{\sum_p Y_p} - \frac{\sum_p M_p (1 - Y_p)}{\sum_p (1 - Y_p)}.
\end{aligned}
\tag{11}
$$

Both quantities are multiplied by 100 in Table 3. Only GT-nonempty cases that generated [SEG] enter these map statistics. We average per case within each dataset and then give the four source or ten OOD datasets equal weight.

Across the registered 3,648 evaluation cases, the complete model produced 3,090 [SEG] and 558 [REJ] responses, all consistent with GT nonempty/empty status. The matched w/o-LMR variant produced one false [REJ] on a nonempty BUID case; that case receives zero DSC and NSD.

A.4 Training and Implementation Details

Table 5: Configuration shared by the four formal LaSE runs.

Item	Setting
Input resolution	512×512
Transformer layers	34
Image tokens	256/layer
spatial map	16×16
MedSAM2 embedding	$256 \times 32 \times 32$
MEGF width	128/branch
LoRA targets	all q, v projections
Rank/alpha/dropout	$8/16/0.05$
Optimizer	AdamW
Weight decay	10^{-4}
LoRA/downstream LR	$3\times10^{-4}/2\times10^{-4}$
LR schedule/warm-up	constant / none
Image augmentation	none beyond deterministic resizing
Gradient clipping	1.0
Global batch	32
Precision	BF16
Maximum epochs	60
Early stopping	patience 12
Selection metric	Val Soft Dice
Decoder calls/case	[SEG]: 1; [REJ]: 1
LMR	enabled

The foundation MedGemma transformer, multimodal projector, vision encoder, and MedSAM2 image encoder are frozen. Query/value LoRA adapters, the tied input/output embedding, MEGF, sparse projection, mask-prompt downscaling, MedSAM2 mask decoder, and LMR are trainable. The four source checkpoints were selected using source validation only.

MEGF scorer. Independent LayerNorm–Linear–GELU projections map the control-token state and candidate-pooled MedSAM2 feature to 128 dimensions each. A shared $256 \to 128 \to 128 \to 1$ MLP with GELU hidden activations scores their concatenation. Its learned temperature is initialized to 1 and constrained to $[0.25, 8]$ before $\mathrm{entmax}_{1.5}$.

Map normalization and training losses. For the layer scores $S_{\ell,p}$ in Eq. (1), LTM uses $M_{\ell,p} = (S_{\ell,p} - \min_j S_{\ell,j})/(\max_j S_{\ell,j} - \min_j S_{\ell,j})$. No epsilon is added to this denominator: non-finite scores or a zero range raise an error rather than producing an imputed map. Let Y be a training mask, A its area-resized 32×32 map, and B_3 a 3×3 mean filter with replicated padding. The soft map target is $Y_s = 0.75A + 0.25B_3(A)$, and the occupancy target Y_o is obtained by adaptive max-pooling of Y to 32×32. For map supervision, let $M_F^{32} = \mathrm{Resize}_{32}(M_F)$ be the bilinearly resized fused map. The auxiliary map objective applies BCE-with-logits directly to this bounded score, soft Dice and boundary losses to $\sigma(M_F^{32})$ against Y_s, and an occupancy-focal loss against Y_o. Their respective weights are 0.25, 0.25, 0.025, and 0.10. This objective trains a spatial prompt; $\sigma(M_F^{32})$ is not treated as a calibrated foreground probability or as the final segmentation mask. Below, $D_{\mathrm{train}}(P, Y)$ denotes per-case soft Dice with a 10^{-5} smoothing constant, averaged across cases.

Let $\widehat{Z}$ denote the *refined* 512×512 mask logits from Eq. (8), and let Y_{512} be the area-resized training mask. The principal mask loss is

$$\mathcal{L}_{\mathrm{mask}} = \mathrm{BCELogits}(\widehat{Z}, Y_{512}) + 1 - D_{\mathrm{train}}(\sigma(\widehat{Z}), Y_{512}) + 0.08\mathcal{L}_{\mathrm{small}}. \tag{12}$$

$\mathcal{L}_{\text{small}}$ averages $1 - (\text{TP} + 10^{-5})/(\text{TP} + 0.30\text{FP} + 0.70\text{FN} + 10^{-5})$ over cases with weight $w = \mathbf{1}[0 < a < 0.01]\,\sigma((0.01 - a)/0.002)$, where a is the target foreground fraction; the weighted sum is divided by $\max(\sum w, 1)$. Only non-empty [SEG] rows enter the map, mask, and LMR losses; empty [REJ] rows contribute language cross-entropy only.

Using the coarse logits Z^{coarse} from Eq. (8), the LMR-only auxiliary loss is

$$\mathcal{L}^{\text{aux}}_{\text{LMR}} = 0.08\mathcal{L}_{\text{coarse}} + 0.08\mathcal{L}_{\text{SDF}} + 0.05\mathcal{L}_{\text{boundary}} + 0.04\mathcal{L}_{\text{scale}} + 0.05\mathcal{L}_{\text{consistency}}. \tag{13}$$

The terms are respectively BCE-with-logits on Z^{coarse}, SmoothL1 on the $\tanh$-transformed signed-distance prediction (averaged over non-empty cases), BCE-with-logits on the boundary prediction, SmoothL1 on the sigmoid scale prediction, and SmoothL1 consistency between 64×64 resized $\sigma(\widehat{Z})$ and detached $\sigma(Z^{\text{coarse}})$. For the area-resized 64×64 target Y_{64}, the SDF target is the outside-minus-inside Euclidean distance divided by 64; the boundary target is the 3×3 max-pool of Y_{64} minus its min-pool, and the scale target is the mean of Y_{64}. The confident-interior indicator is $I = \mathbf{1}[Y_{64} > 0.75]\mathbf{1}[Y_{\text{boundary}} < 0.25]$. The consistency loss is the pixelwise SmoothL1 distance weighted by $I + 0.25(1 - Y_{64})$, divided by the sum of those weights. The language term is assistant-response cross-entropy with weight 1.0. Images are resized bilinearly and masks by nearest-neighbor interpolation; no random geometric or color augmentation is applied.

A.5 Inference Contract

For one non-empty case, replay yields a control trajectory of shape 34×2560 and an image-token trajectory of shape $34 \times 256 \times 2560$. The layer-wise candidate bank has shape $34 \times 1 \times 16 \times 16$, the frozen MedSAM embedding has shape $256 \times 32 \times 32$, and the learned composition is $\boldsymbol{\beta} \in \mathbb{R}^{34}$, producing a fused map $M_F \in \mathbb{R}^{1\times16\times16}$ and a sparse source of dimension 2560.

During training, target masks determine the teacher-forced [SEG]/[REJ] labels and provide map, segmentation, and refinement supervision for non-empty cases. Validation, Test, and OOD evaluation use greedy deterministic generation (`do_sample=false`, one beam, at most 32 new tokens) followed by replay. Target masks are not used in the inference path. A generated [REJ] completes the same decoder call before its output is overwritten with the empty mask.

A.6 Lightweight Mask Refinement

LMR receives Z^{low}, the first low-resolution mask-logit channel from the single MedSAM2 decoder call. Direct interpolation produces Z^{coarse}. A small convolutional pyramid at resolutions 32, 64, 128, and 256 predicts $\Delta_{\text{LMR}}(Z^{\text{low}})$; adding it to Z^{coarse} gives the refined logits $\widehat{Z}$ at 512×512. Auxiliary heads at training time predict a normalized signed-distance field, a boundary map, and lesion scale; the target generator derives an occupancy mask from the training mask, while an interior-consistency term limits corrections in confident regions. At inference, LMR consumes only decoder logits. It contains 29,799 trainable parameters, does not read spatial maps or $\boldsymbol{\beta}$, and does not trigger another mask-decoder call.

A.7 From User Prompt to Spatial Map and Mask

Figure 5 follows four held-out target cases, one from each source family. MedGemma 1.5 generates [SEG] from the supplied image and description; replay exposes the layer-wise control and image states. LTM and MEGF form M_F and the aligned sparse prompt, which guide MedSAM2 decoding before LMR produces the displayed mask.

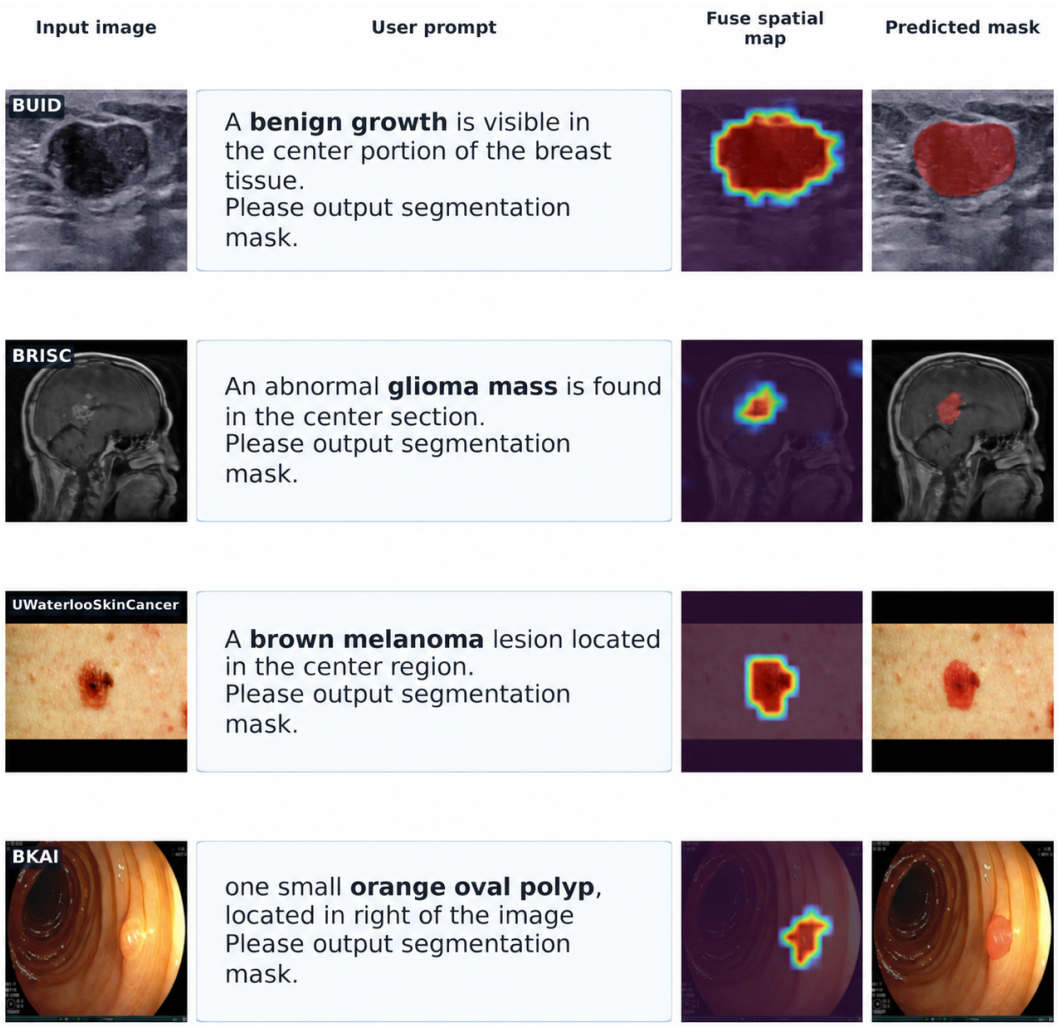


Figure 5: Case studies of LaSE, showing text prompts, fused spatial maps, and final masks for four OOD images.

A.8 Inference Algorithm

Algorithm 1 summarizes the same generate–replay–decode path used for the prompt-to-mask examples.

Algorithm 1 LaSE inference for one image–text pair

Require: image X^I, user prompt X^P, $L = 34$, $N = 256$
Ensure: grounding mask X^G

$R \leftarrow \mathrm{Generate}(X^I, X^P)$
$\{T_\ell^{\mathrm{SEG}}, T_\ell^I\}_{\ell=1}^{L} \leftarrow \mathrm{Replay}(X^I, X^P, R)$
$(F^{\mathrm{SAM}}, F^{\mathrm{hi}}) \leftarrow \xi_{\mathrm{SAM2}}(X^I)$ ▷ one frozen image encoding
$\mathrm{PE} \leftarrow \mathrm{GetDensePE}_{\mathrm{MedSAM2}}()$
for $\ell = 1, \ldots, L$ **do**
  $S_{\ell,p} \leftarrow (T_\ell^{\mathrm{SEG}})^\top t_{\ell,p}^I$ for $p = 1, \ldots, N$
  $M_\ell \leftarrow \mathrm{MinMax}_p(\mathrm{reshape}(S_{\ell,1:N}))$
  $\widetilde{M}_\ell \leftarrow \mathrm{Resize}_{32\times32}(M_\ell)$
  $a_{\ell,p} \leftarrow \widetilde{M}_{\ell,p} / \max(\sum_{p'} \widetilde{M}_{\ell,p'}, \varepsilon)$ ▷ p, p' index the 32×32 feature grid
  $u_\ell \leftarrow \sum_p a_{\ell,p} F_{:,p}^{\mathrm{SAM}}$
  $q_\ell \leftarrow f_{\mathrm{MEGF}}([\phi_{\mathrm{SEG}}(T_\ell^{\mathrm{SEG}}) \| \phi_{\mathrm{med}}(u_\ell)])$
end for
$\boldsymbol{\beta} \leftarrow \mathrm{entmax}_{1.5}((q_1, \ldots, q_L)/\tau)$
$M_F \leftarrow \sum_{\ell=1}^{L} \beta_\ell M_\ell$; $T_F^{\mathrm{SEG}} \leftarrow \sum_{\ell=1}^{L} \beta_\ell T_\ell^{\mathrm{SEG}}$
$D_F \leftarrow \mathcal{P}_{\mathrm{dense}}^{\mathrm{SAM2}}(\mathrm{Resize}(M_F))$; $E_F^{\mathrm{sp}} \leftarrow [e_\varnothing \| \mathcal{P}_{\mathrm{sparse}}^{\mathrm{SEG}}(T_F^{\mathrm{SEG}})]$
$Z^{\mathrm{low}} \leftarrow [\mathcal{D}_{\mathrm{MedSAM2}}(F^{\mathrm{SAM}}, F^{\mathrm{hi}}, \mathrm{PE}, E_F^{\mathrm{sp}}, D_F)]_1$ ▷ first mask-logit channel; one decoder call
$Z^{\mathrm{coarse}} \leftarrow \mathrm{Resize}(Z^{\mathrm{low}})$
$\widehat{Z} \leftarrow Z^{\mathrm{coarse}} + \Delta_{\mathrm{LMR}}(Z^{\mathrm{low}})$
if $[\mathrm{REJ}] \in R$ **then**
  $\widehat{Z} \leftarrow -40$; $X^G \leftarrow \mathbf{0}_{\mathrm{mask}}$
else
  $X^G \leftarrow \sigma(\widehat{Z})$
end if
return X^G

A.9 Training and Inference Complexity

Table 6 reports local BUSI measurements for the benchmark methods and Ariadne's Thread (Zhong et al., 2023). A training step is one optimizer update with an effective global batch size of 32: all contributing microbatches and distributed ranks are included. Training time covers the forward pass, loss computation, backward pass, and optimizer update. LaSE uses four GPUs with eight cases per rank. Inference uses batch size one; after ten warm-up calls, we average the first timed occurrence of each of the 78 distinct BUSI Test cases.

Table 6: Local BUSI training and inference complexity. Training values are per effective B32 optimizer update; inference values are per B1 case. FLOPs are partial MAC-derived arithmetic estimates (1 MAC = 2 FLOPs).

Method	Params (M) Total/Train.	Train Time (s/B32 update)	Train Partial FLOPs (G/B32 update)	Inf. Time (s/B1 case)	Inf. Partial FLOPs (G/B1 case)
LViT	148.825/39.933	0.387	5,200.465	0.0528	57.162
Ariadne's Thread	153.581/43.951	0.212	1,834.813	0.0347	22.973
CLIPSeg	150.743/1.122	0.308	3,228.676	0.0157	98.739
DenseCLIP	113.830/113.830	1.788	12,843.675	0.0422	267.198
ZegCLIP	164.907/14.651	1.424[†]	12,730.359	0.0443	516.195
SAN	157.840/8.371	1.701	11,811.621	0.0330	170.641
CAT-Seg	154.289/25.116	1.775	11,534.997	0.0515	153.225
MedCLIPSeg	214.581/18.677	0.527	6,127.518	0.0493	184.826
LaSE	4,345.108/678.653	2.368	311,939.516	0.7729	15,293.020

Note. [†] ZegCLIP training time is one direct observation; the other training times are means. The LaSE training-FLOPs estimate averages three B32 updates. Total parameters include frozen modules; trainable parameters count unique trainable elements.

The FLOPs columns count supported forward/backward or inference arithmetic, not every operator, optimizer operation, or distributed communication. Input resolution, precision, GPU count, and timing boundaries vary across implementations; these measurements are therefore not a hardware-matched speed ranking. Published MedCLIPSeg Table S2 values (Koleilat et al., 2026) are not mixed with the local measurements.

Formal runs record their software state, source snapshot, configuration, checkpoint, dataset roster, and evaluator. Deterministic algorithms are not forced.

A.10 Additional Qualitative Results

Figures 6–11 extend the ten-target segmentation examples in Figures 3–4. Each figure has five dataset columns and three rows: ground truth, MedCLIPSeg, and LaSE. Figures 12–13 show the ten-target pre-decoder spatial maps.

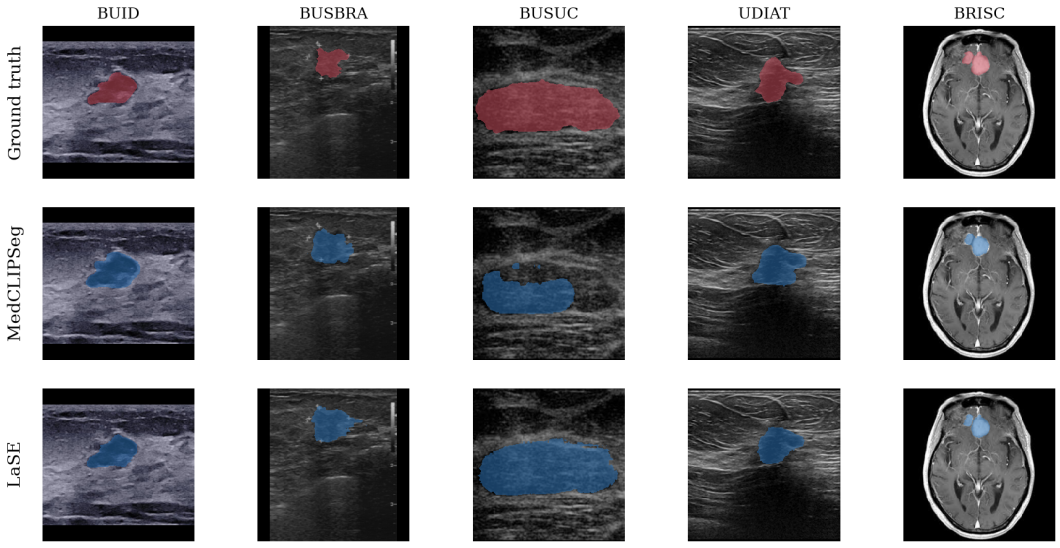


Figure 6: Additional OOD segmentation case studies, set 1: BUID, BUSBRA, BUSUC, UDIAT, and BRISC.

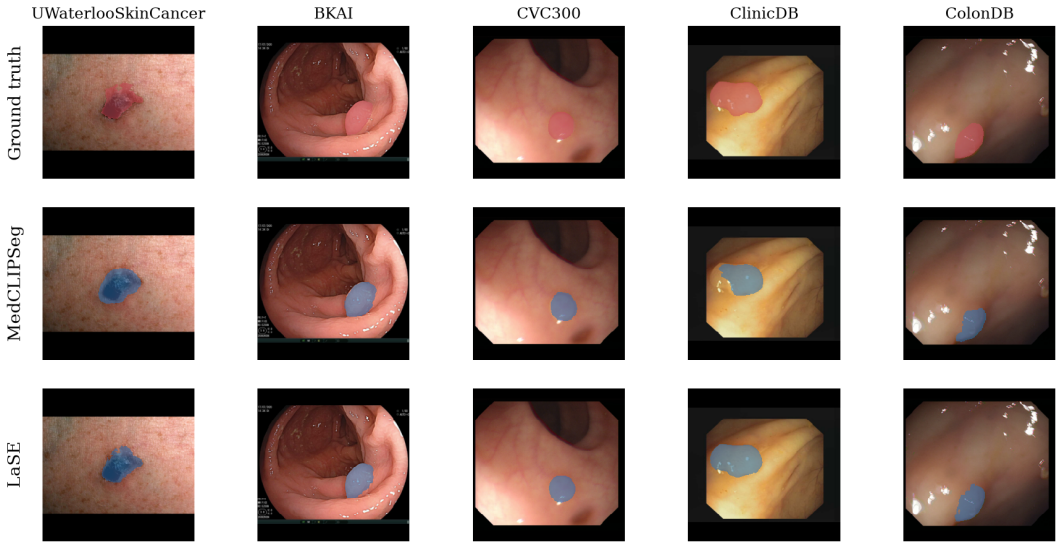


Figure 7: Additional OOD segmentation case studies, set 1: UWaterlooSkinCancer, BKAI, CVC300, ClinicDB, and ColonDB.

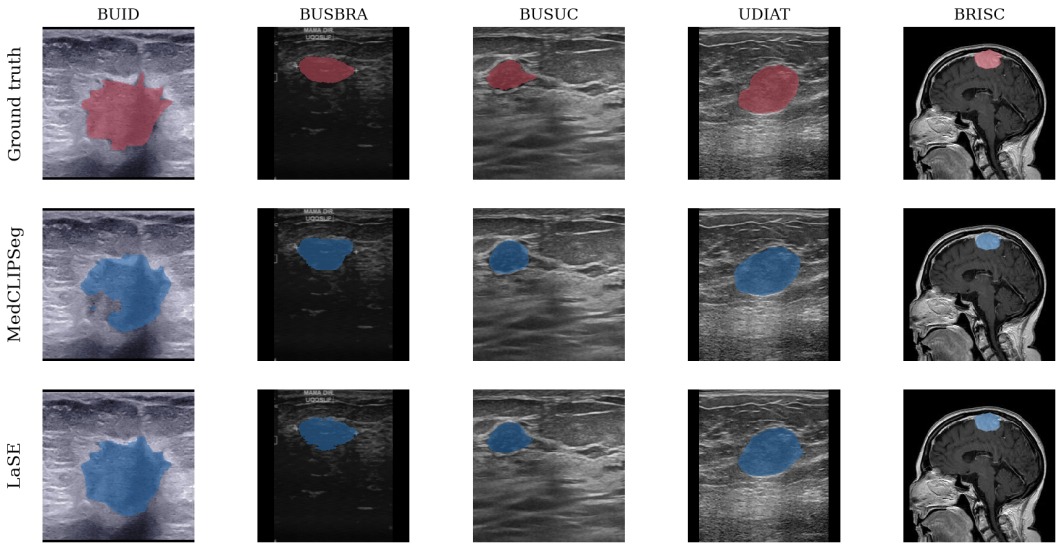


Figure 8: Additional OOD segmentation case studies, set 2: BUID, BUSBRA, BUSUC, UDIAT, and BRISC.

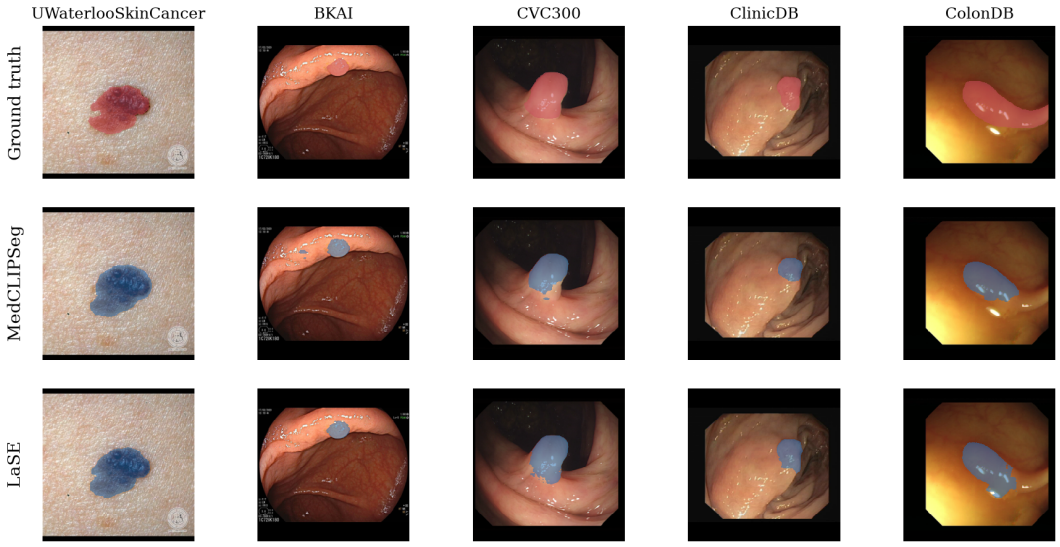


Figure 9: Additional OOD segmentation case studies, set 2: UWaterlooSkinCancer, BKAI, CVC300, ClinicDB, and ColonDB.

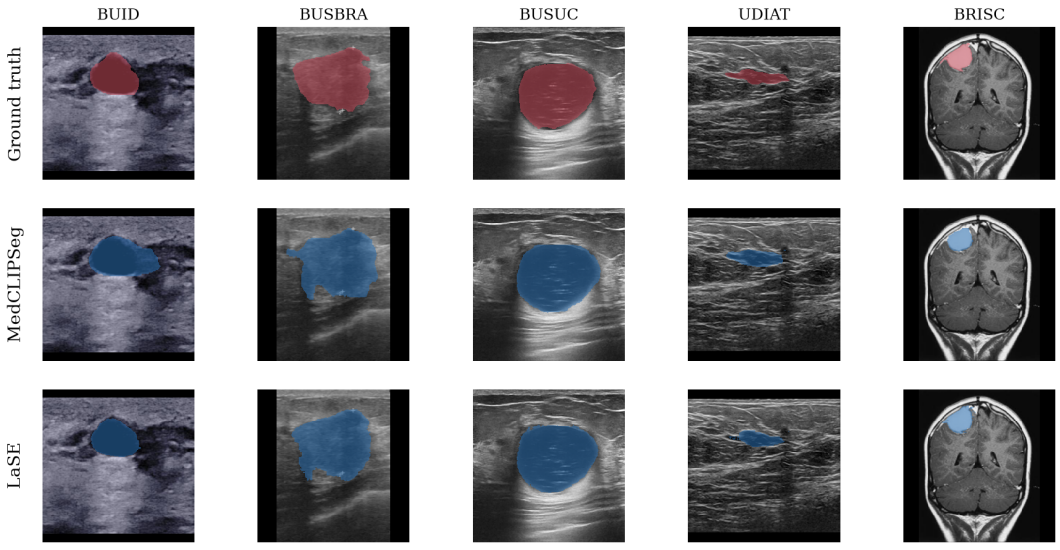


Figure 10: Additional OOD segmentation case studies, set 3: BUID, BUSBRA, BUSUC, UDIAT, and BRISC.

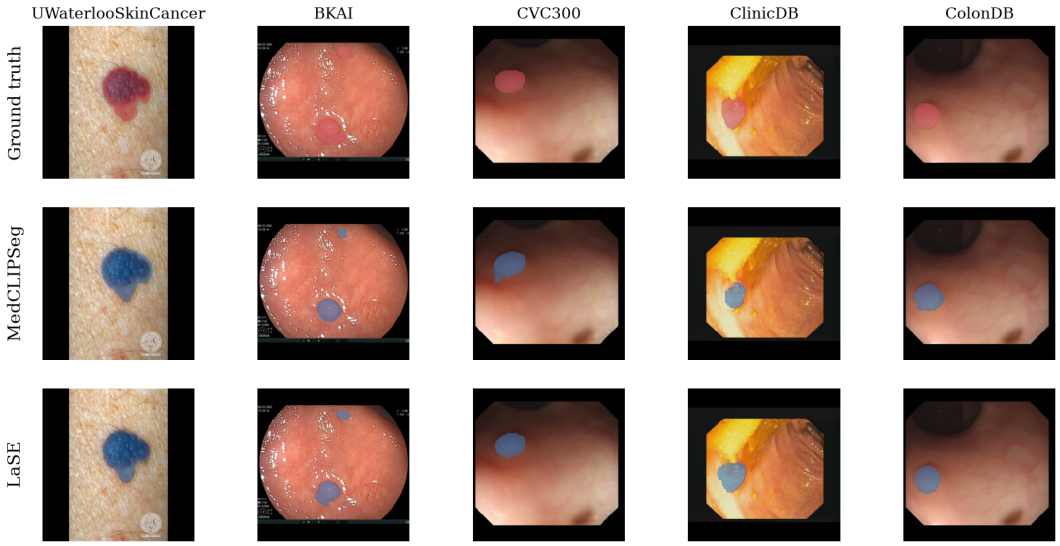


Figure 11: Additional OOD segmentation case studies, set 3: UWaterlooSkinCancer, BKAI, CVC300, ClinicDB, and ColonDB.

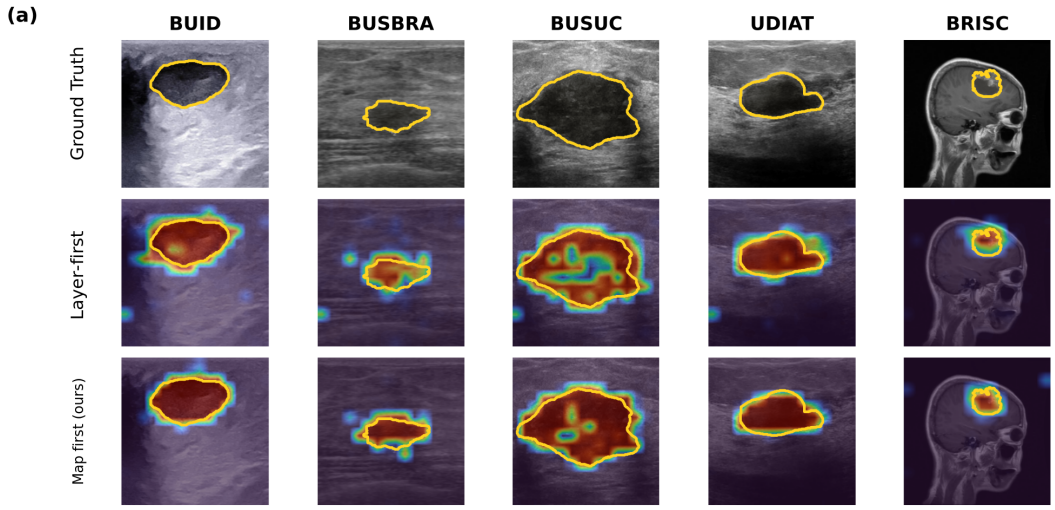


Figure 12: Similarity-map case studies on BUID, BUSBRA, BUSUC, UDIAT, and BRISC, comparing layer-first routing with LaSE.

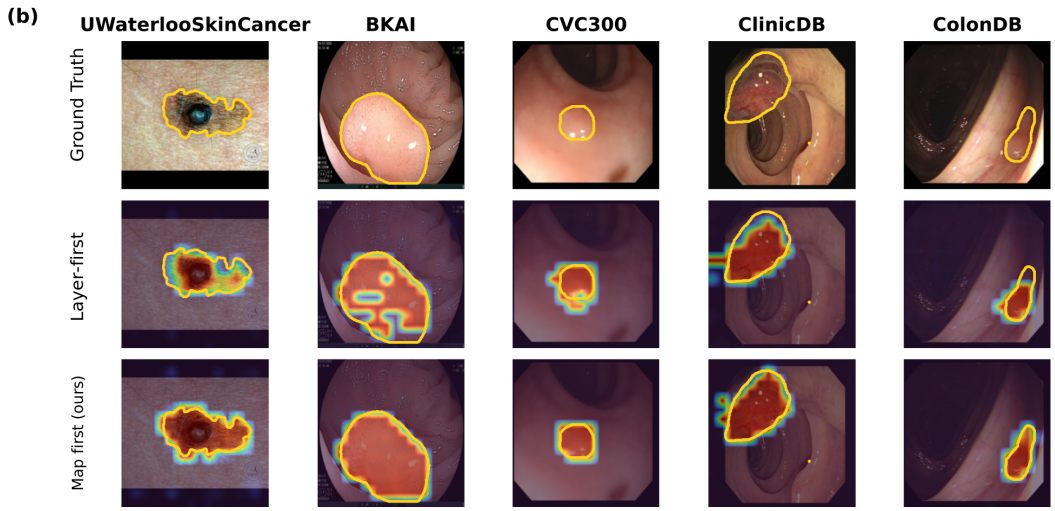


Figure 13: Similarity-map case studies on UWaterlooSkinCancer, BKAI, CVC300, ClinicDB, and ColonDB, comparing layer-first routing with LaSE.

A.11 Layer-Weight Distribution

The learned β assigns continuous, sparse weights across the 34 spatial maps. Figure 14 presents paired source and OOD bars for each source family. BUSI concentrates on layers 21–22, BTMRI on 19–20, ISIC on 25–31, and Kvasir on 25–28. The bar heights also show how the average composition changes from source cases to their corresponding OOD targets, especially for BTMRI and Kvasir. Each OOD bar is an equal-target mean after within-target case averaging.

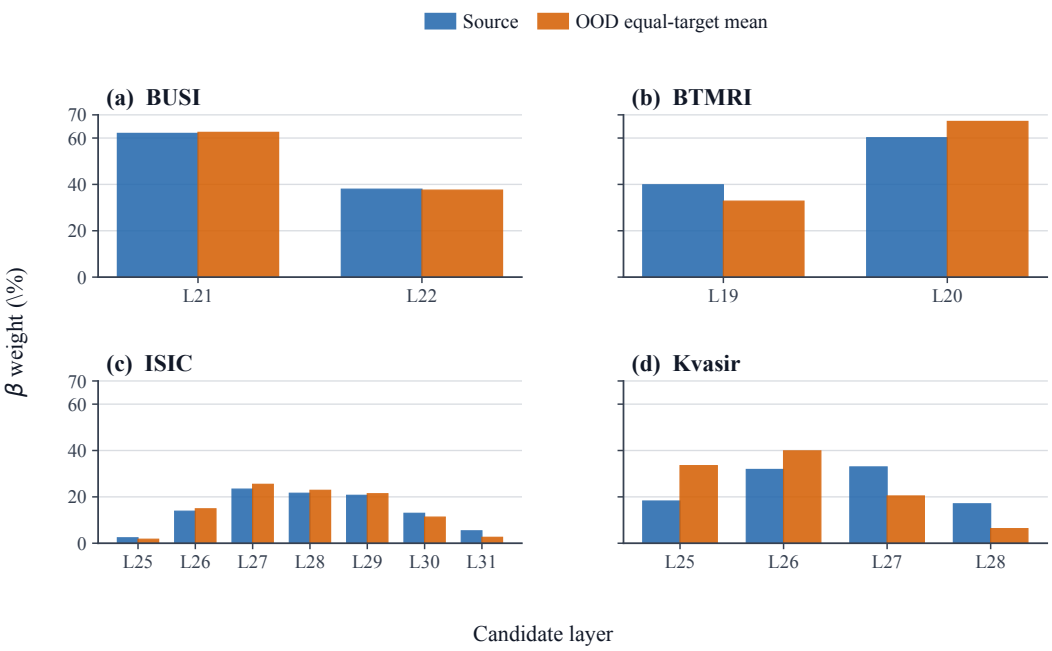


Figure 14: Layer-wise β weights for each source family. Blue and orange bars are source-case and equal-target OOD means over 3,090 valid [SEG] cases. All panels share a 0–70% axis. Displayed layers exceed 0.5 percentage points in at least one group; omitted layers jointly contribute less than 0.1% in either group.