SGFA: SPATIALLY GUIDED FEATURE ADAPTATION FOR MULTI-DISEASE ULTRASOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultrasound lesions from different diseases can have similar intensity and geometry but differ in local texture, making ambiguous boundaries difficult to resolve from spatial cues alone. We propose SGFA (Spatially Guided Feature Adaptation), a multi-disease ultrasound lesion segmentation framework combining Shared Texture Projection (STP) with Quadrant-wise Boundary Box Prompting (QBBP). STP generates image-dependent residual feature updates through projection parameters shared across disease categories. QBBP encodes quadrant-wise boundary boxes as sparse and dense prompts and constrains these updates to prompt-supported regions through a spatial gate. We evaluate SGFA using ground-truth-derived (oracle) prompts on a five-disease pancreatic endoscopic ultrasound (EUS) test set comprising 8,531 frames from 164 patients. We additionally evaluate the framework on the MMOTU ovarian ultrasound dataset with dataset-specific training. SGFA achieves a patient-mean Dice of 0.9548 (frame-mean: 0.9563) on EUS and an image-mean Dice of 0.9642 on MMOTU. Patient-level EUS ablations show gains from STP adaptation under global-box prompting and from jointly adapting STP and the spatial gate under QBBP.

1 INTRODUCTION

Ultrasound lesion segmentation requires a representation that accommodates heterogeneous tissue appearances while preserving weak and irregular boundaries. This challenge is particularly relevant to pancreatic endoscopic ultrasound (EUS), an established setting for evaluating pancreatic lesions (Machicado et al., 2024). Across diseases, lesions may occupy similar spatial regions yet exhibit different local intensity patterns and texture. A geometric prompt can identify the region of interest, but delineating the lesion also requires interpreting the image features within that region. Multi-disease EUS therefore motivates a joint consideration of appearance adaptation and spatial localization.

Progress in medical segmentation has established effective convolutional and transformer representations (Ronneberger et al., 2015; Isensee et al., 2021; Hatamizadeh et al., 2022). More recently, Segment Anything (SAM) has introduced a reusable interface for point- and box-conditioned mask prediction (Kirillov et al., 2023), with medical and ultrasound adaptations extending its applicability (Ma et al., 2024; Lin et al., 2024; Li et al., 2026). In parallel, Anatomy-aware Representation Learning (ARL) emphasizes ultrasound-specific feature learning through anatomy-adaptive pretraining (Oh et al., 2026). These developments motivate examining how a segmentation representation can accommodate local appearance variation across diseases within the same anatomical setting.

We study this question at the interface between image features and spatial prompts. Feature adaptation determines the information presented to the mask decoder, whereas prompting specifies spatial constraints on the target. Their combination offers a direct way to refine appearance-sensitive features within a localized region. The central question is whether a compact, shared projection contributes to multi-disease lesion delineation beyond the contribution of spatial prompting itself.

We propose SGFA, combining Shared Texture Projection (STP) with Quadrant-wise Boundary Box Prompting (QBBP). STP transforms the final image embedding through parallel local operators and a low-rank bottleneck. Its parameters are shared across disease categories, while its residual response varies with each image. QBBP encodes four local boxes as sparse corner tokens and dense spatial priors, and constrains residual writeback through a support-restricted gate. This division of roles

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.

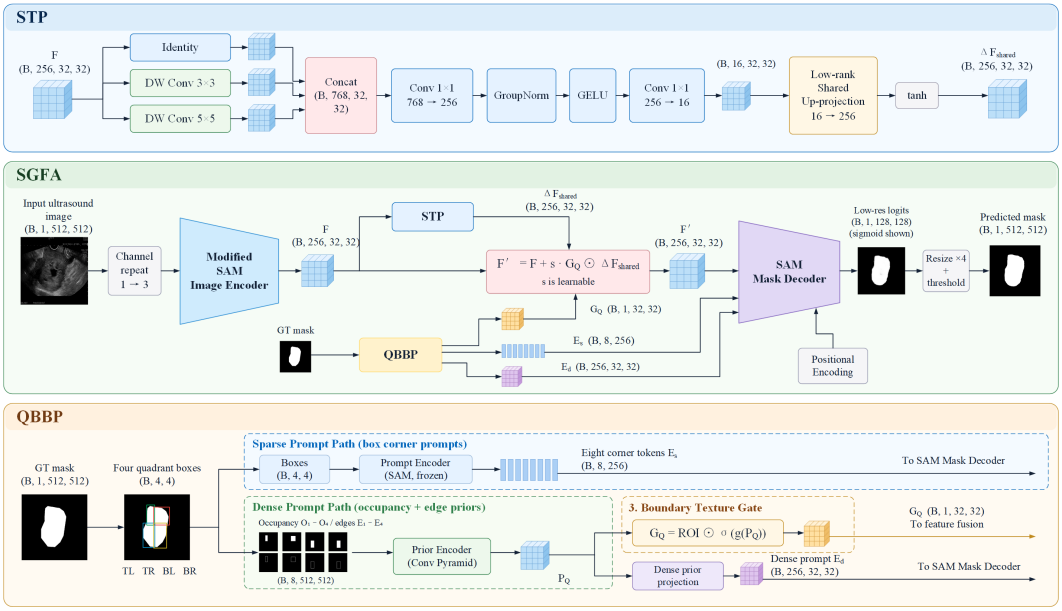

Figure 1: SGFA architecture and its two modules; tensor dimensions illustrate the EUS configuration. Top: STP aggregates the image embedding through parallel spatial operators and a shared projection to produce a bounded residual. Middle: QBBP-gated writeback adapts the image features for mask decoding. Bottom: GT-derived quadrant boxes provide sparse corner tokens, dense occupancy/edge priors, and a support-restricted spatial gate.

separates the feature update from its spatial application: STP adapts the representation, while QBBP determines where the update acts. The resulting features and prompts are processed in a single mask-decoder call.

Five-disease pancreatic EUS provides the core multi-disease evidence, with matched component comparisons and equal-weight patient-level analysis. STP adaptation increases patient-mean Dice by 0.003120 with a global box; joint adaptation of STP and the spatial gate increases it by 0.003175 with QBBP, improving 133 of 164 patients. MMOTU provides complementary cross-anatomy validation through dataset-specific training and testing on ovarian ultrasound, where SGFA reaches a Dice of 0.964159. Prompt-information and capacity controls further characterize the spatial interface and feature-adaptation pathway.

Our contributions are threefold: (i) a shared, image-dependent feature projection for multi-disease lesion segmentation; (ii) a quadrant-based prompt representation that couples decoder conditioning with spatially constrained residual writeback; and (iii) patient-level evidence for projection adaptation in five-disease EUS, complemented by cross-anatomy validation and controlled analyses on MMOTU.

2 Related Work

Medical segmentation and ultrasound representations. Dense prediction architectures have progressed from fully convolutional networks and U-Net to multiscale and transformer-based systems (Long et al., 2015; Ronneberger et al., 2015; Chen et al., 2018; Hatamizadeh et al., 2022). Volumetric modeling, overlap-aware objectives, attention, and self-configuring pipelines further support medical segmentation (Milletari et al., 2016; Sudre et al., 2017; Oktay et al., 2018; Isensee et al., 2021). DANet and SegFormer model spatial context without requiring inference-time prompts (Fu et al., 2019; Xie et al., 2021), while UniverSeg and SegVol explore support-conditioned and volumetric segmentation interfaces (Butoi et al., 2023; Du et al., 2024). For ultrasound, ARL learns anatomy-aware representations using an anatomy-adaptive transformer and self-supervised pretraining (Oh et al., 2026). Our focus is complementary: adapting the final embedding of a trained segmentation backbone to heterogeneous lesion appearances, with shared projection parameters and explicit spatial constraints.

Promptable medical and ultrasound segmentation. SAM and SAM 2 establish promptable segmentation interfaces for images and videos (Kirillov et al., 2023; Ravi et al., 2025). MedSAM and SAM-Med2D adapt SAM to medical imagery (Ma et al., 2024; Cheng et al., 2023); medical evaluations and few-shot studies characterize its transfer across downstream tasks (Roy et al., 2023; Xie et al., 2024). HQ-SAM and H-SAM develop mask-quality and hierarchical-decoding improvements (Ke et al., 2023; Cheng et al., 2024). Ultrasound-specific approaches include SAMUS, four-point-prompted S4M, and the SAM 2-based UniUltra (Lin et al., 2024; Meyer et al., 2026; Li et al., 2026). These methods provide the specialized comparison group for our evaluation. SGFA examines a complementary design axis: coupling a shared feature residual with a spatially localized prompting representation.

Parameter-efficient representation adaptation. Low-rank updates and lightweight adapters provide compact alternatives to full-model fine-tuning. LoRA parameterizes weight updates with low-rank factors (Hu et al., 2022), while AdaptFormer introduces trainable modules for adapting frozen vision transformers (Chen et al., 2022a). SAM-Adapter and SAM-DA extend adapter-based specialization to segmentation (Chen et al., 2023; Tejero et al., 2025). STP operates on the final image embedding rather than updating backbone weights: local feature aggregation and a bottleneck produce a bounded residual, whose writeback is constrained by the prompt. Our component and parameter-matched comparisons evaluate this feature-adaptation branch within a fixed segmentation interface.

Spatial prompting and boundary modeling. Extreme points and sparse annotations encode object geometry compactly (Maninis et al., 2018; En & Guo, 2024; Shaharabany & Wolf, 2025), while interactive methods refine masks from user clicks (Sofiiuk et al., 2022; Chen et al., 2022b; Liu et al., 2023). PointRend and Mask2Former improve dense prediction, and Boundary IoU emphasizes boundary-sensitive evaluation (Kirillov et al., 2020; Cheng et al., 2022; 2021). Curve-GCN and Deep Snake instead represent and update ordered contours (Ling et al., 2019; Peng et al., 2020). QBBP encodes quadrant-local box geometry as sparse tokens and dense occupancy/edge priors before decoding. The same geometry constrains feature writeback, connecting spatial prompting with representation adaptation in a single prediction pass.

3 METHOD

3.1 SHARED FEATURE ADAPTATION WITH SPATIAL CONSTRAINTS

Let $I \in \mathbb{R}^{1\times H\times W}$ be an ultrasound image and $\mathcal{B} = \{b_q\}_{q=1}^{4}$ its quadrant-box prompt. The segmentation target is a binary lesion mask. Images are resized to 512×512 and repeated across three input channels. A SAMUS-initialized image encoder produces

$$F = E(I) \in \mathbb{R}^{256\times 32\times 32}. \tag{1}$$

STP operates on this final embedding, outside the internal encoder blocks. QBBP generates sparse tokens, a dense prompt, and a spatial gate. Their outputs meet at a single mask decoder (Figure 1). Batch dimensions are omitted for clarity.

3.2 SHARED TEXTURE PROJECTION (STP)

STP shares its projection parameters across disease categories while computing a residual from each image embedding. The branch combines an identity path with depthwise 3×3 and 5×5 convolutions:

$$T = \phi\big([F, \mathrm{DWConv}_3(F), \mathrm{DWConv}_5(F)]\big), \qquad T \in \mathbb{R}^{256\times 32\times 32}, \tag{2}$$

where concatenation gives 768 channels and ϕ contains a 1×1 fusion, group normalization, and GELU. A rank-r bottleneck ($r = 16$ for EUS and $r = 8$ for MMOTU) maps this representation to a bounded residual:

$$Z = W_{\downarrow}T, \qquad \Delta F = \tanh(W_{\uparrow}Z), \qquad Z \in \mathbb{R}^{r\times 32\times 32}. \tag{3}$$

The parallel operators aggregate complementary local context before projection, while the low-rank bottleneck provides a compact residual adaptation path. The projection parameters are shared across disease categories, enabling the same branch to operate across heterogeneous lesion appearances without introducing disease-specific parameters.

3.3 QUADRANT-WISE BOUNDARY BOX PROMPTING (QBBP)

Reference contour points are partitioned around their arithmetic mean into four quadrants. Each box $b_q = (x_q^0, y_q^0, x_q^1, y_q^1)$ encloses its quadrant points and this center, with per-side padding of 5% (at least one input pixel; Appendix A). Two corner embeddings per box produce $E_s \in \mathbb{R}^{8\times 256}$. The same coordinates yield occupancy and edge-response maps:

$$O_q(x) = \mathbf{1}[x \in b_q], \qquad A_q(x) = O_q(x) \exp\left[-\frac{d(x, \partial b_q)^2}{2\sigma_e^2}\right]. \tag{4}$$

Invalid boxes are masked out. These maps encode box occupancy and proximity to box edges. An eight-channel stack is passed through the prior encoder:

$$P_Q = \mathrm{PriorEncoder}([O_1, \ldots, O_4, A_1, \ldots, A_4]), \qquad E_d = E_{\text{no-mask}} + \lambda \tanh(W_P P_Q). \tag{5}$$

P_Q and E_d have the same spatial dimensions and channel count as F. Four stride-two stages reduce the maps from 512^2 to 32^2, followed by an embedding-resolution convolution.

QBBP separates spatial support from the image-dependent residual: the encoded prior determines where STP's update is written back, through a learned gate restricted to the quadrant-box support:

$$G_Q = \mathrm{sigmoid}(g(P_Q)) \odot \mathrm{Downsample}_{\text{nearest}}\left(\max_q O_q\right), \qquad F' = F + sG_Q \odot \Delta F. \tag{6}$$

The gate has one channel and is broadcast across the embedding channels. STP thus determines the residual features, while QBBP limits their direct writeback to prompt-supported locations. For the global-box variant, the gate reduces to the downsampled single-box support, E_s contains two corner tokens, and $E_d = E_{\text{no-mask}}$. Without STP, $F' = F$. Finally,

$$L = D(F', E_s, E_d), \qquad \widehat{M} = \mathrm{Upsample}(L), \tag{7}$$

where L has resolution 128^2 and the output logits have resolution 512^2. SGFA predicts the segmentation mask in a single decoder pass. Throughout this study, QBBP is a *GT-derived prompted-segmentation protocol*.

3.4 STAGED FEATURE ADAPTATION

We evaluate the global-box baseline, STP, QBBP only, and SGFA, which combines STP and QBBP. Adaptation starts from the corresponding validation-selected source checkpoint and runs for 3,000 steps, with validation Dice selecting the final checkpoint. The image encoder, prompt encoder, mask decoder, and dense-prior pathway remain frozen; STP and, with QBBP, the spatial gate are learned jointly. On MMOTU, the adaptation pathway contains 357,570 parameters: 209,921 in STP, including its residual scale, and 147,649 in the spatial gate. The corresponding pathway contains 361,666 parameters on EUS, which uses a rank-16 bottleneck. Layerwise counts and optimizer parameter groups are detailed in Appendix A.1.

This design measures the incremental value of feature adaptation under each spatial interface. In MMOTU, all four variants use seeds 42, 3407, and 1234, with each projection branch initialized from its seed-matched reference checkpoint. In the EUS paired patient analysis, the global-box and QBBP-only models provide fixed references on a patient-isolated split. Each reference supports three independent adaptation runs using the same seed set and is evaluated once. STP is trained with lesion-mask supervision and shared parameters across disease categories.

4 EXPERIMENTS

4.1 DATA, BASELINES, AND REPORTING PROTOCOL

EUS: core multi-disease cohort. The private pancreatic EUS test cohort contains 8,531 frames from 164 patients, isolated from development patients. Its five categories are autoimmune pancreatitis (AIP; 10 patients, 1,135 frames), pancreatic cancer (CA; 116, 5,184), chronic pancreatitis (CP; 9, 389), neuroendocrine tumor (NET; 23, 1,341), and solid pseudopapillary tumor (SPT; 6, 482). These are held-out test counts, with each patient assigned to one category. Patient-level aggregation gives each patient equal weight despite differing frame counts. The retrospective study was approved by the institutional ethics review board of the participating hospital. The requirement for informed consent was waived, and all images were de-identified before analysis.

MMOTU: validation on ovarian ultrasound. MMOTU OTU_2D contains 1,469 ovarian ultrasound images (Zhao et al., 2023). We retain the official 469-image test set and use a fixed 800/200 training/validation split of the remaining images. This experiment evaluates the applicability of the framework beyond pancreatic EUS through dataset-specific training and held-out testing.

Baselines and selection. We compare SegFormer-B5, DANet-R101, SAM, MedSAM, SAMUS, S4M, and UniUltra, with the same baseline implementations used across EUS and MMOTU. Table 1 identifies each method's source publication. Each baseline uses its native inference interface, with prompt types differing across methods as listed in Table 4. S4M uses the official 1,600-iteration schedule. MMOTU uses seeds 42, 3407, and 1234. Checkpoints are selected exclusively by validation Dice. Matched component and capacity controls are evaluated separately from the main comparison.

Metrics and uncertainty. We report Dice, IoU, boundary F1 (BF1), and the 95th-percentile Hausdorff distance (HD95), capturing both overlap and boundary agreement (Cheng et al., 2021; Karimi & Salcudean, 2020). All predictions were resampled to the native ground-truth resolution before metric computation. Binary masks use nearest-neighbor resampling; scores are computed per image before averaging. BF1@2 is the primary boundary metric for both EUS and MMOTU, and HD95 is measured in native-image pixels. Table 1 reports mean $\pm$ sample SD across three training seeds for every method on both datasets. Within each seed, metrics are averaged over EUS frames or MMOTU images before computing the across-seed summary. The patient-level EUS comparison averages frames within each patient and then weights the 164 patients equally. All four EUS component variants report frame-level mean $\pm$ sample SD across seeds 42, 3407, and 1234. The paired patient analysis uses the matched frozen source checkpoints; its patient-level estimates are reported separately from the frame-level summaries.

For paired MMOTU contrasts, we first average each image's metric difference across matched seeds, then bootstrap the 469 image-level differences 20,000 times. EUS uses 20,000 paired resamples of the 164 patient-level differences. Paired bootstrap percentile 95% confidence intervals (CIs) quantify case-sampling uncertainty conditional on the evaluated models. Variation across training seeds is reported separately as the sample SD. Appendix A gives the aggregation definitions.

4.2 Comparison with existing methods

SGFA achieves the highest mean Dice and BF1@2 and the lowest HD95 among the evaluated methods on both datasets (Table 1). We first examine five-disease EUS, then assess whether the same framework remains effective in ovarian ultrasound.

Multi-disease EUS performance. SGFA reaches frame Dice 0.95629, BF1@2 0.85091, and HD95 6.2421 pixels. Relative to S4M, the strongest comparator on all three EUS metrics, Dice increases by 4.08 percentage points and BF1@2 by 13.52 points, while HD95 decreases by 6.03 pixels. The improvement therefore covers both lesion overlap and boundary agreement. Boundary-sensitive evaluation is informative here: the relative ordering of several baselines differs between Dice and BF1, whereas SGFA leads on both measures.

The patient-level results preserve this pattern when each patient receives equal weight (Appendix C). SGFA obtains Dice 0.95475, BF1@2 0.84331, and HD95 6.3639 pixels; S4M reaches 0.9122, 0.6975, and 12.2683, respectively. SGFA's patient and frame Dice differ by only 0.15 percentage points, indicating that its high cohort-level overlap is retained under equal-patient aggregation. This complements the frame-weighted comparison in a cohort with unequal numbers of images per patient.

Cross-anatomy performance on MMOTU. SGFA attains Dice 0.9642 and IoU 0.9315, exceeding S4M by approximately 0.74 and 1.30 percentage points. The strongest baseline depends on the boundary measure: MedSAM has the highest baseline BF1@2 (0.4497), whereas S4M has the lowest baseline HD95 (12.5181 pixels). SGFA improves these values to 0.5062 and 11.9675, respectively. Thus, the regional-overlap advantage is accompanied by better boundary metrics in a second anatomical setting. Dataset-specific training and held-out testing establish the applicability of the architecture to ovarian ultrasound, complementing the multi-disease EUS evidence.

Figure 2 complements the EUS benchmark with examples from all five disease categories.

Table 1: Segmentation performance on EUS and MMOTU.

Method	Ref.	Five-disease EUS			MMOTU		
		Dice ↑	BF1@2 ↑	HD95 ↓	Dice ↑	BF1@2 ↑	HD95 ↓
General-purpose segmentation baselines							
SegFormer-B5	Xie et al. (2021)	.86277±.00316	.63749±.00433	15.6324±.1598	.8726±.0022	.2941±.0061	41.4739±.7036
DANet-R101	Fu et al. (2019)	.86394±.00056	.64786±.00132	15.7943±.2589	.8586±.0027	.2558±.0119	47.7744±.2590
SAM-family baselines							
SAM	Kirillov et al. (2023)	.8673±.0968	.5362±.2170	17.6625±15.9608	.9420±.0006	.3722±.0019	18.9004±.2337
MedSAM	Ma et al. (2024)	.8697±.0949	.5417±.2165	17.3115±15.2903	.9544±.0011	.4497±.0088	14.6697±.8904
Ultrasound-specialized models and ours							
S4M	Meyer et al. (2026)	.9155±.0734	.7157±.2070	12.2685±13.4061	.9568±.0003	.4250±.0023	12.5181±.4431
SAMUS	Lin et al. (2024)	.8960±.0842	.6399±.2273	14.4738±14.6331	.8427±.0081	.2273±.0114	46.9179±1.7608
UniUltra	Li et al. (2026)	.7706±.1781	.4260±.2290	41.5634±44.7001	.8640±.0005	.2782±.0012	41.4223±.5885
SGFA	Ours	**.95629**±.00006	**.85091**±.00026	**6.2421**±.0531	**.9642**±.0004	**.5062**±.0030	**11.9675**±.3198

Native-protocol comparison. All values are mean ± sample SD across three training seeds. Within each seed, EUS scores are frame-weighted and MMOTU scores are image-averaged. Bold/underline: best/second-best mean. HD95 is in pixels. Patient-level EUS results: Appendix C.

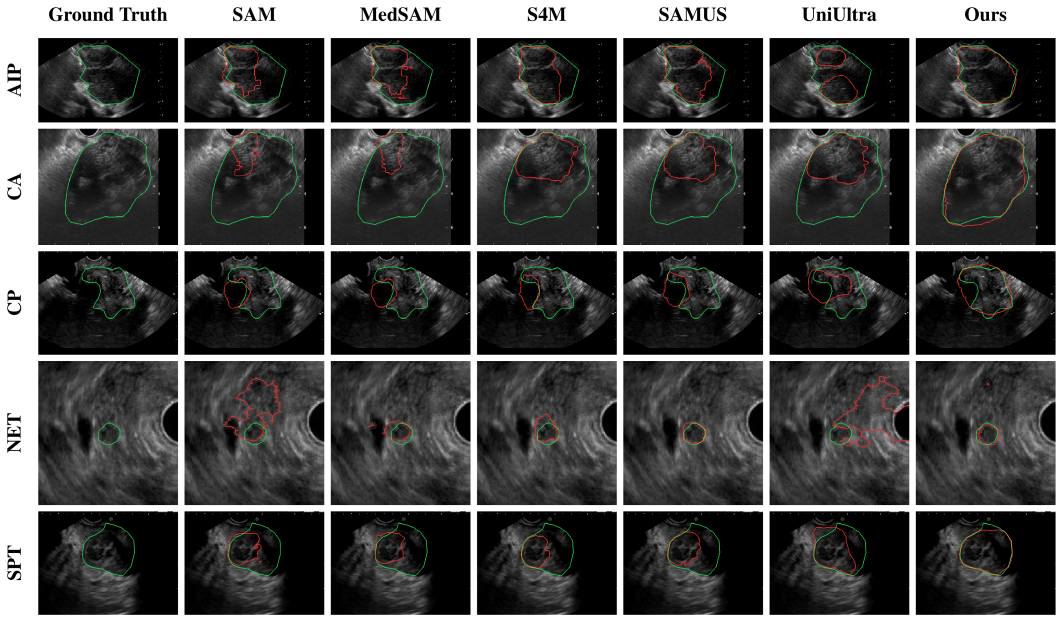

Figure 2: Five-disease EUS comparator examples, ordered by AIP, CA, CP, NET, and SPT. Columns show Ground Truth, SAM, MedSAM, S4M, SAMUS, UniUltra, and Ours. Green: reference contour; red: prediction. The same crop is applied to all methods within each row to remove peripheral acquisition text.

4.3 Feature-adaptation evidence in multi-disease EUS

Patient-level adaptation effects on EUS. The paired patient analysis compares adaptation with its matched frozen source under each spatial interface in the five-disease EUS cohort (Table 2). Under global-box prompting, STP adaptation increases patient-mean Dice by 0.003120 (95% CI $[0.001949, 0.004376]$; 106/164 patients improved). Under QBBP, jointly adapting STP and the spatial gate increases patient-mean Dice by 0.003175 ($[0.002442, 0.003954]$; 133/164 improved). This joint adaptation also increases patient BF1@2 by 0.011518 ($[0.006750, 0.016230]$) and decreases HD95

by 0.653769 pixels ($[-0.878466, -0.466274]$). Feature adaptation improves patient-level Dice under both prompting interfaces, with additional boundary-quality gains observed under QBBP.

Table 2: STP and QBBP ablations on EUS and MMOTU.

		Five-disease EUS					MMOTU			
STP	QBBP	Frame Dice ↑	Frame IoU ↑	Patient Dice ↑	BF1@2 ↑	HD95 ↓	Dice ↑	IoU ↑	BF1@2 ↑	HD95 ↓
–	–	.9427 ±.0017	.8931 ±.0030	.941072	.6923 ±.0142	8.1901 ±.2032	.947178 ±.000789	.901731 ±.001473	.386832 ±.006713	16.755120 ±.414931
✓	–	.946168 ±.000085	.899336 ±.000136	.944192 ±.000117	.809295 ±.000198	7.786276 ±.029087	.948690 ±.000736	.904364 ±.001343	.396300 ±.004448	16.339302 ±.375895
–	✓	.9536 ±.0027	.9121 ±.0050	.951573	.7521 ±.0194	6.6366 ±.0565	.963576 ±.000491	.930426 ±.000895	.500391 ±.005301	12.202690 ±.332870
✓	✓	**.956289** ±.000063	**.916933** ±.000103	**.954748** ±.000160	**.850912** ±.000260	**6.242102** ±.053113	**.964159** ±.000398	**.931493** ±.000726	**.506171** ±.003000	**11.967460** ±.319757

QBBP off uses a global box; both modules on gives SGFA, with STP and the spatial gate adapted jointly. SDs appear below three-seed means. EUS frame metrics and patient Dice use frame and equal-patient weighting, respectively. Patient Dice comes from the matched-source analysis (fixed references have no seed SD). Boundary metrics use native resolution.

The frame-weighted EUS summaries in Table 2 also favor adaptation under both spatial interfaces. From the reported means, STP adaptation raises Dice by approximately 0.35 percentage points with a global box, while joint STP–gate adaptation raises it by 0.27 with QBBP. Relative to QBBP only, SGFA increases BF1@2 by approximately 0.0988 and reduces HD95 by 0.3945 pixels. These descriptive differences complement the matched-checkpoint patient-level effects and intervals in Appendix A.

Cross-anatomy validation on MMOTU. The MMOTU columns of Table 2 show the same direction of adaptation effects on ovarian ultrasound. STP adaptation improves Dice by 0.001513 with a global box (95% CI $[0.001218, 0.001822]$); joint STP–gate adaptation improves it by 0.000583 with QBBP ($[0.000449, 0.000721]$). Relative to QBBP only, SGFA improves 297/469 images, increases BF1@2 by 0.005780, and reduces HD95 by 0.235230 pixels. QBBP alone improves Dice by 0.016398 over the global-box baseline, while SGFA exceeds STP with a global box by 0.015469 ($[0.013658, 0.017462]$). These adaptation gains extend to a second anatomical setting under dataset-specific training.

4.4 Prompt-information controls

Prompt geometry beyond the number of cues. Table 3(a) compares four prompt interfaces with the SAMUS backbone, training budget, and test split held fixed across seeds 42, 3407, and 1234. Four quadrant-center points are derived from the same boxes used by QBBP. QBBP improves Dice, IoU, and BF1@2 and reduces HD95 relative to every control in all three seeds. Against four quadrant centers, the paired gains are 0.009513 Dice and 0.118369 BF1@2, with HD95 reduced by 1.618798 pixels. Against four axis points, Dice and BF1@2 increase by 0.008448 and 0.075872, respectively.

The two four-point controls hold the number of localized cues fixed but encode different aspects of lesion geometry: axis endpoints specify extent along major directions, while quadrant centers specify local position. QBBP provides local spatial extent through eight corner tokens and a dense prior. Its advantage over both controls supports the joint value of structured boxes and dense spatial conditioning, beyond the number of prompted regions alone. Table 3(b) summarizes the seed-paired differences.

Table 3: Prompt controls on MMOTU.

Setting / reference	Dice ↑	IoU ↑	BF1@2 ↑	HD95 ↓
(a) Prompt information				
1 GT tight box	.947479 ±.002100	.902272 ±.003493	.392706 ±.006756	16.838520 ±1.053969
4 GT major/minor-axis points	.955128 ±.000448	.915263 ±.000790	.424519 ±.005061	14.276473 ±.202559
4 quadrant-center points	.954062 ±.001027	.913145 ±.001839	.382022 ±.004614	13.821488 ±.311427
4 QBBP boxes + dense prior	**.963576** ±.000491	**.930426** ±.000895	**.500391** ±.005301	**12.202690** ±.332870
(b) Paired differences: QBBP minus each control				
versus 1 GT tight box	+.016097 ±.002551	+.028154 ±.004301	+.107684 ±.011485	−4.635830 ±1.168463
versus 4 axis points	+.008448 ±.000939	+.015163 ±.001681	+.075872 ±.009539	−2.073784 ±.511829
versus 4 quadrant centers	+.009513 ±.001020	+.017281 ±.001784	+.118369 ±.006992	−1.618798 ±.083344

Means with sample SDs below ($n = 3$ seeds; 469 test images). Panel (b) reports seed-paired QBBP-minus-control differences. BF1 tolerance is 2 native pixels; HD95 is in pixels.

Appendix B reports the pointwise-adapter control under the same QBBP interface, including per-metric results and paired intervals.

Sensitivity to prompt accuracy. For the frozen QBBP-only seed-42 model, Dice changes from 0.963758 with clean boxes to 0.955818, 0.940060, and 0.886547 under 5%, 10%, and 20% coordinate jitter; removing one quadrant gives 0.828496. Performance decreases with increasing coordinate jitter and incomplete quadrant coverage. Perturbation definitions and boundary metrics are provided in Appendix B.

4.5 MMOTU QUALITATIVE ANALYSIS AND COMPUTATIONAL COST

Figure 3 shows the largest positive and negative seed-averaged Dice changes for SGFA relative to QBBP only. In the displayed improvement, Dice rises from approximately 0.795 to 0.803; in the displayed regression, it changes from approximately 0.978 to 0.973. These examples show how joint STP–gate adaptation modifies an already localized prediction. Across the full test set, 297/469 images improve and 172 decline. Full qualitative panels are provided in Appendix D.

Table 4 separates matched-512^2 measurements from methods evaluated at their native internal grids. On the matched grid, SGFA requires 24.13 ms and $1{,}668$ MiB per image, with 324.11 GFLOPs. Its higher segmentation accuracy is obtained with greater compute and memory than SAM or MedSAM; latency remains below SegFormer-B5 in this measurement. Native-grid results retain each method's architecture and input resolution.

The matched component profile in Appendix Table 8 measures the incremental cost of the STP-and-gate pathway: SGFA uses 0.369 more GFLOPs and 1.132 ms more latency than QBBP only, with essentially unchanged peak memory. Measurement procedures are given in Appendix A.

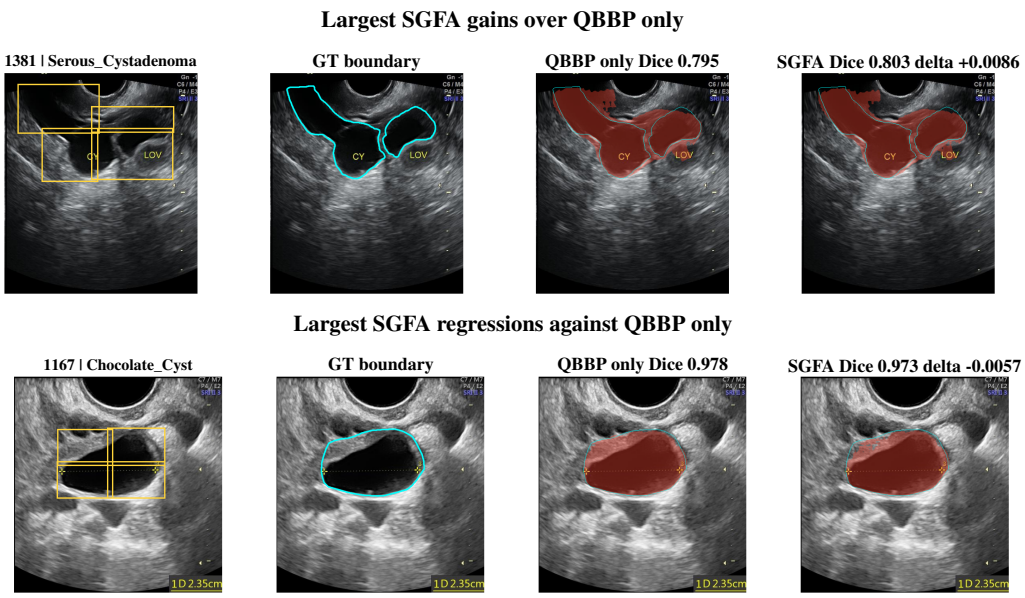

Figure 3: Cases with the largest seed-averaged Dice gain (top) and loss (bottom) for SGFA relative to QBBP only on the MMOTU test set. Columns: QBBP-prompted input, GT, QBBP-only prediction, and SGFA prediction. Full panels: Appendix D.

Table 4: Computational efficiency of the compared methods.

Method	Input	Prompt	Params (M)	GFLOPs	Latency (ms)	Peak MiB	Dice ↑
(a) Strict matched-input comparison							
SegFormer-B5	512^2	None	84.59	219.27	38.58	1129	.8726±.0022
DANet-R101	512^2	None	66.43	574.08	9.85	563	.8586±.0027
SAM	512^2	GT box	91.34	228.03	15.28	494	.9420±.0006
MedSAM	512^2	GT box	91.34	228.03	15.14	653	.9544±.0011
SGFA	512^2	4 GT quadrant boxes	273.96	324.11	24.13	1668	**.9642**±.0004
(b) Evaluation at native input resolutions							
SAMUS	256^2	1 GT point	133.32	326.30	19.53	748	.8427±.0081
S4M	1024 pipeline	4 GT major/minor points	97.24	N/A	353.04	2557	.9568±.0003
UniUltra	1024^2	GT component boxes	233.53	N/A	47.77	1279	.8640±.0005

NVIDIA A800-SXM4-80GB; BF16; batch size one; frozen seed 42. Panels (a)/(b) use matched/native input grids. Dice is mean ± SD across three seeds. N/A: unsupported FLOP coverage.

5 Conclusion

SGFA addresses heterogeneous lesion appearance through a projection shared across disease categories and spatially constrained residual writeback. STP produces image-dependent feature updates, while QBBP defines their spatial support and supplies prompts to the mask decoder. Five-disease EUS provides patient-level evidence for STP adaptation under global-box prompting and joint STP–gate adaptation under QBBP. MMOTU extends the evaluation to ovarian ultrasound under dataset-specific training. Together, these results support combining shared feature adaptation and localized spatial constraints for GT-derived prompted segmentation.

AI USE STATEMENT

Generative AI tools assisted language editing, manuscript organization, drafting, and code-related suggestions. Experimental findings derive from model evaluations and statistical analyses. The authors retain responsibility for verifying all AI-assisted content and for the final manuscript.

REFERENCES

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 21438–21451, 2023.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pp. 801–818. Springer, 2018.

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*, 2022a.

Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. SAM-Adapter: Adapting segment anything in underperformed scenes. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3367–3375, 2023.

Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1300–1309, 2022b.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15334–15342, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. SAM-Med2D. *arXiv preprint arXiv:2308.16184*, 2023.

Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, and Yuyin Zhou. Unleashing the potential of SAM for medical adaptation via hierarchical decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3511–3522, 2024.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Qing En and Yuhong Guo. Annotation by clicks: A point-supervised contrastive variance method for medical semantic segmentation. In *35th British Machine Vision Conference*. BMVA, 2024. URL [link].

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, 2019.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3d medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.

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.

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(2): 203–211, 2021.

Davood Karimi and Septimiu E. Salcudean. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 39(2):499–513, 2020.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9799–9808, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Yue Li, Qing Xu, Yixuan Zhang, Xiangjian He, Qian Zhang, Yuan Yao, Fiseha Berhanu Tesem, Xin Chen, Ruili Wang, Zhen Chen, and Chang Wen Chen. UniUltra: Interactive parameter-efficient SAM2 for universal ultrasound segmentation. *IEEE Transactions on Multimedia*, 2026.

Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008 of *Lecture Notes in Computer Science*, pp. 24–34. Springer Nature Switzerland, 2024.

Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with Curve-GCN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5257–5266, 2019.

Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 22290–22300, 2023.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

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

Jorge D. Machicado, Sunil G. Sheth, Jean M. Chalhoub, Nauzer Forbes, Madhav Desai, et al. American society for gastrointestinal endoscopy guideline on the role of endoscopy in the diagnosis and management of solid pancreatic masses: Summary and recommendations. *Gastrointestinal Endoscopy*, 100(5):786–796, 2024.

Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 616–625, 2018.

Adrien Meyer, Lorenzo Arboit, Giuseppe Massimiani, Shih-Min Yin, Didier Mutter, and Nicolas Padoy. S4M: 4-points to segment anything. *International Journal of Computer Assisted Radiology and Surgery*, 21(6): 1311–1319, 2026.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision*, pp. 565–571, 2016.

Seok-Hwan Oh, Myeong-Gee Kim, Guil Jung, Hyeon-Jik Lee, Young-Min Kim, Sang-Yun Kim, Hyuksool Kwon, and Hyeon-Min Bae. Anatomy-aware representation learning for medical ultrasound. In *International Conference on Learning Representations*, 2026.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Sida Peng, Wen Jiang, Huajin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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 Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL [link].

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015.

Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R. Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H. Maier-Hein. SAM.MD: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023.

Tal Shaharabany and Lior Wolf. WS-SAM: Segment anything with sparse prompts for weakly-supervised medical segmentation. In *21st International Symposium on Medical Information Processing and Analysis*, pp. 1–5. IEEE, 2025.

Konstantin Sofiiuk, Ilia A. Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *IEEE International Conference on Image Processing*, pp. 3141–3145, 2022.

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, volume 10553 of *Lecture Notes in Computer Science*, pp. 240–248. Springer, 2017.

Javier Gamazo Tejero, Moritz J. Schmid, Pablo Márquez-Neila, Martin Zinkernagel, Sebastian Wolf, and Raphael Sznitman. SAM-DA: Decoder adapter for efficient medical domain adaptation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6775–6784, 2025.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Weiyi Xie, Nathalie Willems, Shubham Patil, Yang Li, and Mayank Kumar. SAM fewshot finetuning for anatomical segmentation in medical images. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3253–3261, 2024.

Qi Zhao, Shuchang Lyu, Wenpei Bai, Linghan Cai, Binghao Liu, Guangliang Cheng, Meijing Wu, Xiubo Sang, Min Yang, and Lijiang Chen. MMOTU: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. *arXiv preprint arXiv:2207.06799*, 2023. URL [link].