TIGR-SEG: TEXT-INITIALIZED GAUSSIAN REFINEMENT WITH SPATIALLY AWARE VISUAL ADAPTATION FOR MEDICAL IMAGE SEGMENTATION

Anonymous Authors
Paper under Double-blind Review

ABSTRACT

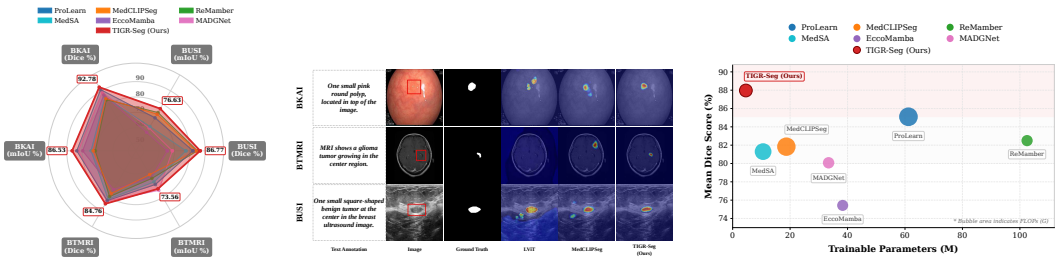
Medical image segmentation remains challenging due to the limited availability of pixel-level annotations, heterogeneous target appearance, and variations across imaging modalities. Pretrained vision encoders provide strong representations, but updating all backbone parameters introduces substantial trainable capacity, while textual information often provides semantic and coarse spatial guidance rather than precise pixel-level localization. This work presents TIGR-Seg, a parameter-efficient text-guided segmentation framework comprising Gaussian-Gated Feature Adaptation (GGFA) and a Text-Initialized Gaussian Refinement (TIGR) Decoder. GGFA adapts a frozen pretrained vision encoder through context-conditioned feature modulation and a dynamically predicted correlated Gaussian spatial gate, enabling adaptation across both feature and spatial dimensions while updating only a small set of parameters. The TIGR Decoder incorporates stage-specific textual guidance across multiple decoding resolutions. At each stage, text initializes a Gaussian spatial hypothesis that is subsequently refined using visual evidence through reliability-weighted corrections. The refined spatial priors then guide feature aggregation, reconstruction, and progressive decoder refinement. Experiments across gastrointestinal endoscopy (BKAI), brain MRI (BTMRI), and breast ultrasound (BUSI) datasets demonstrate strong segmentation performance with a low trainable-parameter count. Additional evaluations support the effectiveness and generalizability of TIGR-Seg.

1 INTRODUCTION

Medical image segmentation is a fundamental task in medical image analysis and plays an important role in disease diagnosis, treatment planning, and quantitative clinical analysis (Azad et al., 2024a). Despite substantial advances in convolutional and transformer-based architectures (Azad et al., 2024b), reliable segmentation remains challenging because pathological regions can vary considerably in size, morphology, texture, contrast, and spatial location (Litjens et al., 2017). These challenges are further influenced by modality-specific characteristics, as illustrated in Figure 1(b): endoscopic images can exhibit illumination variations and visually similar surrounding structures (Qayoom et al., 2025), magnetic resonance imaging (MRI) can present heterogeneous lesion appearance and weak tissue contrast (Ranjbarzadeh et al., 2023), while ultrasound images are often affected by speckle noise and acoustic artifacts, making accurate target localization particularly difficult for small or visually ambiguous abnormalities (Xian et al., 2018). Moreover, obtaining high-quality pixel-level annotations requires substantial time and clinical expertise, limiting the availability of densely annotated medical datasets and the amount of task-specific supervision available for model training (Conze et al., 2023).

Clinical reports and structured descriptions can provide clinically relevant information such as target anatomy, abnormality type, lesion count, approximate location, and coarse appearance (Larson et al., 2013). Such textual information is treated as an auxiliary input modality that provides complementary semantic and coarse spatial guidance rather than pixel-level boundary information (Lu & Wang, 2025), (Li et al., 2025). Text-guided segmentation methods, including LViT (Li et al., 2023), BiomedParse (Zhao et al., 2025), TeViA (Zeng et al., 2025), C2Seg (Gui et al., 2026),

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.



(a) Radar Plot (b) Qualitative Comparison (c) Accuracy-Efficiency Comparison

Figure 1: **Quantitative, qualitative, and efficiency analysis of TIGR-Seg. (a)** Radar plot comparing Dice and mIoU on BKAI (endoscopy) (Ngoc Lan et al., 2021), BTMRI (brain MRI) (Cheng, 2017), and BUSI (breast ultrasound) (Al-Dhabyani et al., 2020). **(b)** Qualitative comparison showing textual annotations, input images, ground-truth masks, and Grad-CAM visualizations for LViT, MedCLIPSeg, and TIGR-Seg; red boxes indicate the target regions for visualization only. **(c)** Accuracy-efficiency comparison, where the horizontal axis denotes trainable parameters, the vertical axis represents the mean Dice score across the three datasets, and bubble area indicates FLOPs.

HiMix (Hwang et al., 2026), and MedCLIPSeg (Koleilat et al., 2026), further incorporate such textual information into the segmentation process.

Yet two key challenges remain in text-guided medical image segmentation. First, updating all parameters of a pretrained vision backbone introduces substantial trainable capacity and can increase the risk of overfitting when task-specific medical data are limited. Parameter-efficient fine-tuning (PEFT) (Dutt et al., 2024) reduces this burden by limiting the number of trainable parameters; however, dense segmentation can further benefit from adaptation that accounts for the spatial structure of visual features. Second, textual and visual information differ in spatial precision. Textual descriptions provide semantic information and coarse spatial context, whereas accurate segmentation requires detailed localization from the image. Such textual guidance can also be incomplete or only partially aligned with the visual evidence. Therefore, rather than using text as fixed localization information, its spatial guidance should be verified and refined using image evidence before contributing to pixel-level segmentation.

To address these challenges, TIGR-Seg introduces Gaussian-Gated Feature Adaptation (GGFA) and a Text-Initialized Gaussian Refinement (TIGR) Decoder. GGFA enables parameter-efficient adaptation of a frozen pretrained vision encoder by combining context-conditioned feature modulation with a dynamically predicted Gaussian spatial gate, accounting for both feature content and spatial structure. The TIGR Decoder incorporates stage-specific textual guidance across multiple decoding resolutions, where text initializes a spatial hypothesis that is subsequently refined using visual evidence through reliability-weighted corrections. The refined spatial information then guides progressive feature decoding. TIGR-Seg thus integrates spatially aware visual adaptation with image-refined textual guidance within a unified segmentation framework. Fig. 1 provides a comparative overview of segmentation performance, qualitative localization, and computational efficiency. The main contributions of this work are summarized as follows:

- GGFA provides a parameter-efficient mechanism for adapting a frozen pretrained vision encoder by combining context-conditioned feature modulation with a dynamic correlated Gaussian spatial gate, enabling adaptation across both feature and spatial dimensions with a small number of trainable parameters.
- The TIGR Decoder formulates textual guidance as an initial spatial hypothesis that is refined using visual evidence. Reliability-weighted image corrections progressively update the text-initialized spatial information before the resulting Gaussian priors guide multi-scale decoder refinement.
- Comprehensive experiments across endoscopy, MRI, and ultrasound datasets validate TIGR-Seg in terms of segmentation accuracy and computational efficiency. Further analyses assess the contributions of the proposed components, textual guidance, reduced-data performance, and cross-dataset generalization.

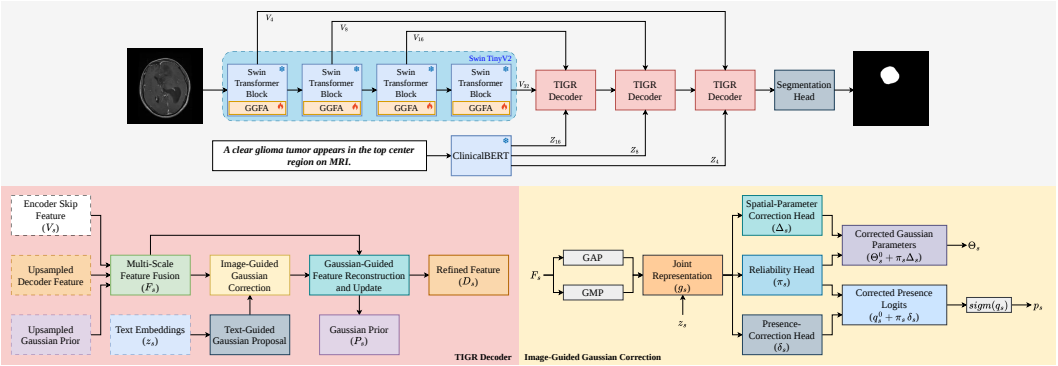


Figure 2: **Overview of TIGR-Seg.** The overall framework is shown together with detailed views of the TIGR Decoder and Image-Guided Gaussian Correction module. GGFA adapts the frozen vision encoder, while the TIGR Decoder progressively refines text-initialized spatial guidance using visual evidence.

2 METHOD

Figure 2 illustrates the overall architecture of TIGR-Seg, which consists of three main components: (i) Dual-Path Encoders that extract hierarchical visual features using Swin-V2-Tiny (Liu et al., 2022) and stage-specific text embeddings derived from a frozen ClinicalBERT encoder (Liu et al., 2025); (ii) Gaussian-Gated Feature Adaptation (GGFA), which selectively adapts the frozen visual backbone using context-conditioned modulation and a dynamic Gaussian gate; and (iii) the Text-Initialized Gaussian Refinement (TIGR) Decoder, which initializes a Gaussian spatial hypothesis from text, refines it using image evidence, and uses the resulting prior to guide feature reconstruction and decoding. The finest decoder feature is mapped to the final segmentation mask. The following subsections describe these components in detail.

2.1 DUAL-PATH ENCODERS

Image Encoder. The image encoder uses a frozen Swin-V2-Tiny (Liu et al., 2022) backbone to extract hierarchical visual features from an input image $I \in \mathbb{R}^{B\times 3\times H\times W}$, where B denotes the batch size and H and W denote the image height and width, respectively:

$$\{V_4, V_8, V_{16}, V_{32}\} = \mathrm{Swin}(I). \tag{1}$$

The feature map at output stride s is denoted by $V_s \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$, where C_s denotes the number of feature channels, $s \in \{4, 8, 16, 32\}$, and $H_s = H/s$ and $W_s = W/s$ are the corresponding spatial dimensions. The channel dimensions are $(C_4, C_8, C_{16}, C_{32}) = (96, 192, 384, 768)$. The coarsest feature V_{32} initializes the decoder, while V_{16}, V_8, and V_4 provide skip features. GGFA provides trainable adaptation within the query, key, and value projections, while the remaining pretrained Swin parameters are frozen.

Text Encoder. The text encoder uses a frozen ClinicalBERT (Liu et al., 2025) encoder to obtain contextual representations from a clinical description $T \in \mathbb{N}^{B\times L}$ with attention mask $m \in \{0, 1\}^{B\times L}$,

$$E = \mathrm{ClinicalBERT}(T, m) \in \mathbb{R}^{B\times L\times d_t}, \tag{2}$$

where $d_t = 768$. The contextual representations are averaged over the non-padding positions, and the resulting shared representation is projected to each decoder stage,

$$\bar{e}_T = \frac{\sum_{\ell=1}^{L} m_\ell E_\ell}{\sum_{\ell=1}^{L} m_\ell}, \qquad z_s = \tau_s(\bar{e}_T). \tag{3}$$

Here, E_ℓ denotes the contextual representation at token position ℓ, and $\bar{e}_T \in \mathbb{R}^{B\times d_t}$ is the shared text representation. The learnable projection τ_s, defined for $s \in \{16, 8, 4\}$, maps $\bar{e}_T$ to $z_s \in \mathbb{R}^{B\times C_s}$, which provides the text embedding for the corresponding decoder stage.

2.2 GAUSSIAN-GATED FEATURE ADAPTATION

GGFA adapts each frozen attention projection through two complementary branches, (i) context-conditioned feature modulation, which determines how the pretrained response is modified, and (ii) dynamic Gaussian gating, which determines where and to what extent the correction is applied within the attention window.

For a Swin attention window of $M \times M$ tokens, $N_w = M^2$ denotes the number of tokens in the window. Let B_w denote the number of windows in a batch, and let $X \in \mathbb{R}^{B_w \times N_w \times d_{\text{in}}}$ denote the input to a frozen query, key, or value projection. The pretrained response is

$$Y_0 = XW_0^\top + b_0, \qquad Y_0 \in \mathbb{R}^{B_w \times N_w \times d_{\text{out}}}, \tag{4}$$

where $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and $b_0 \in \mathbb{R}^{d_{\text{out}}}$ remain fixed during training.

Both branches are conditioned on a shared r-dimensional representation of the current window,

$$h = \text{GELU}\left(\text{LN}\left(\frac{1}{N_w}\sum_{i=1}^{N_w} X_i W_d\right)\right), \qquad h \in \mathbb{R}^{B_w \times r}, \tag{5}$$

where $W_d \in \mathbb{R}^{d_{\text{in}} \times r}$ and $r \ll d_{\text{in}}$. The window-specific representation h provides the shared conditioning signal for both feature modulation and Gaussian gating.

The context-conditioned modulation branch maps h to channel-wise scale and shift parameters,

$$\gamma = \tanh(hW_\gamma + b_\gamma), \qquad \beta = \tanh(hW_\beta + b_\beta), \tag{6}$$

where $W_\gamma, W_\beta \in \mathbb{R}^{r \times d_{\text{out}}}$. After broadcasting γ and β across the N_w window tokens, the affine feature correction is

$$Y_{\text{mod}} = \gamma \odot Y_0 + \beta, \qquad Y_{\text{mod}} \in \mathbb{R}^{B_w \times N_w \times d_{\text{out}}}. \tag{7}$$

The dynamic Gaussian-gating branch maps the shared context h through a two-layer MLP f_G with hidden dimension r to a six-dimensional intermediate parameter vector for each attention window,

$$[\tilde{\mu}_x, \tilde{\mu}_y, \tilde{\sigma}_x, \tilde{\sigma}_y, \tilde{\rho}, \tilde{a}] = f_G(h) \in \mathbb{R}^{B_w \times 6}, \tag{8}$$

which is transformed to the Gaussian-gate parameters as

$$\begin{aligned} \mu_x &= \text{sigm}(\tilde{\mu}_x), \qquad \mu_y = \text{sigm}(\tilde{\mu}_y), \\ \sigma_x &= \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\,\text{sigm}(\tilde{\sigma}_x), \\ \sigma_y &= \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\,\text{sigm}(\tilde{\sigma}_y), \\ \rho &= \rho_{\max}\tanh(\tilde{\rho}), \qquad a = \text{sigm}(\tilde{a}), \end{aligned} \tag{9}$$

where $\text{sigm}(\cdot)$ denotes the logistic sigmoid. The centre $\mu = (\mu_x, \mu_y)$ is defined in normalized window coordinates, σ_x and σ_y determine the spatial extent of the gate, ρ controls the correlation between the two spatial axes, and a controls the gate amplitude. The corresponding covariance matrix is

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\,\sigma_x\sigma_y \\ \rho\,\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}. \tag{10}$$

The parameterization in Eq. (9), with $\sigma_{\min} > 0$ and $\rho_{\max} < 1$, ensures $\sigma_x, \sigma_y > 0$ and $|\rho| < 1$, making Σ positive definite. For the normalized spatial coordinate $t_i \in [0,1]^2$ of token i, the Gaussian gate value is

$$G_i = a \exp\left[-\frac{1}{2}(t_i - \mu)^\top \Sigma^{-1}(t_i - \mu)\right]. \tag{11}$$

Evaluating the Gaussian at all N_w token coordinates yields the token-wise gate values $\{G_i\}_{i=1}^{N_w}$, which form the spatial gate $G \in \mathbb{R}^{B_w \times N_w \times 1}$. The dependence of the Gaussian parameters on the window representation h allows the location, spatial extent, correlation structure, and amplitude of the gate to vary with the content of each attention window.

The two branches are combined as

$$Y = Y_0 + \lambda\, G \odot Y_{\text{mod}}, \tag{12}$$

where G is broadcast along the channel dimension and λ is a learnable residual strength. Here, Y_{mod} determines the feature-space correction, while G controls its spatial contribution within the attention window. The modulation heads are zero-initialized, giving $\gamma = \beta = 0$ and $Y_{\mathrm{mod}} = 0$ at initialization. Consequently, $Y = Y_0$, preserving the pretrained projection at the start of training.

For $d_{\mathrm{in}} = d_{\mathrm{out}} = d$ and $r \ll d$, GGFA requires $\mathcal{O}(rd)$ trainable parameters per adapted projection, compared with $\mathcal{O}(d^2)$ for full fine-tuning of the same projection. The detailed parameter-count analysis and PyTorch-style pseudocode are provided in Supplementary Sec. B.

2.3 Text-Initialized Gaussian Refinement Decoder

The Text-Initialized Gaussian Refinement (TIGR) Decoder progressively refines the visual representation at output strides $s \in \{16, 8, 4\}$. At each stage, the decoder feature from the preceding coarser level is fused with the corresponding encoder skip feature. The stage-specific text embedding initializes a Gaussian spatial hypothesis, which is subsequently refined using visual evidence. The resulting Gaussian prior guides feature aggregation, reconstruction, and the residual decoder update.

The decoder is initialized with $D_{32} = V_{32}$ and $P_{32} = 0$, where D_s and P_s denote the decoder feature and Gaussian prior at output stride s, respectively.

Multi-Scale Feature Fusion. At decoder stage s, the preceding coarser feature D_{2s} is upsampled and fused with the corresponding encoder skip feature V_s,

$$F_s^{(0)} = \phi_s([\mathrm{Up}(D_{2s}); V_s])\,. \tag{13}$$

The preceding Gaussian prior P_{2s} is similarly upsampled and incorporated as spatial context,

$$F_s = \psi_s\Big(\Big[F_s^{(0)}; \mathrm{Up}(P_{2s})\Big]\Big)\,. \tag{14}$$

Here, $F_s^{(0)}, F_s \in \mathbb{R}^{B\times C_s\times H_s\times W_s}$, and $[\cdot;\cdot]$ denotes channel-wise concatenation. The operators ϕ_s and ψ_s share the same architecture but have separate learnable parameters. Each comprises a 1×1 convolution, group normalization, GELU, and a depthwise-separable 3×3 refinement block, combining channel projection with local spatial refinement.

Text-Guided Gaussian Proposal. For each decoder stage s, the stage-specific text embedding $z_s \in \mathbb{R}^{B\times C_s}$ is mapped by a Gaussian-parameter head to the initial spatial parameters Θ_s^0 and by a presence head to the initial target-presence logit q_s^0,

$$\begin{aligned} \Theta_s^0 &= z_s W_{G,s}^{\top} + b_{G,s} \in \mathbb{R}^{B\times 5},\\ q_s^0 &= z_s w_{p,s} + b_{p,s} \in \mathbb{R}^{B\times 1}. \end{aligned} \tag{15}$$

Here, $W_{G,s} \in \mathbb{R}^{5\times C_s}$ and $b_{G,s} \in \mathbb{R}^5$ denote the weight and bias of the Gaussian-parameter head, while $w_{p,s} \in \mathbb{R}^{C_s\times 1}$ and $b_{p,s} \in \mathbb{R}$ denote the weight and bias of the presence head. The text-initialized parameter vector is

$$\Theta_s^0 = \big[\tilde{\mu}_{x,s}^0, \tilde{\mu}_{y,s}^0, \tilde{\sigma}_{x,s}^0, \tilde{\sigma}_{y,s}^0, \tilde{\rho}_s^0\big]\,. \tag{16}$$

The first two components are assigned to the horizontal and vertical centre parameters, the next two to the corresponding spatial-scale parameters, and the final component to the correlation parameter. These spatial parameters are subsequently refined using visual evidence before being mapped to the bounded parameters of the correlated Gaussian.

Image-Guided Gaussian Correction. To refine the text-initialized Gaussian hypothesis, global average and maximum pooling are applied to the fused feature F_s. The resulting visual descriptors are combined with the stage-specific text embedding to form a joint image-text representation

$$g_s = \big[\mathrm{GAP}(F_s); \mathrm{GMP}(F_s); z_s\big] \in \mathbb{R}^{B\times 3C_s}. \tag{17}$$

This representation is processed by three stage-specific heads: a spatial-parameter correction head $f_{\Delta,s}$, a reliability head $f_{\pi,s}$, and a presence-correction head $f_{\delta,s}$. The three heads are implemented as separate two-layer MLPs with GELU activation.

The spatial-parameter correction head predicts a bounded displacement of the text-initialized spatial parameters

$$\Delta_s = \tanh\big(f_{\Delta,s}(g_s)\big) \odot c \in \mathbb{R}^{B\times 5}, \tag{18}$$

where c is a fixed five-dimensional correction-bound vector with positive entries that specifies the maximum displacement for each spatial parameter. The reliability head predicts an input-dependent correction weight

$$\pi_s = \mathrm{sigm}\big(f_{\pi,s}(g_s)\big) \in (0,1)^{B\times 1}. \tag{19}$$

The text-initialized spatial parameters are refined as

$$\Theta_s = \Theta_s^0 + \pi_s \Delta_s \in \mathbb{R}^{B\times 5}, \tag{20}$$

where π_s is broadcast across the five parameter dimensions. Since $|\tanh(\cdot)| \le 1$, the refinement is bounded by $|\Theta_s - \Theta_s^0| \le \pi_s c$. The reliability weight π_s therefore controls the magnitude of the image-guided refinement while constraining its deviation from the text-initialized spatial parameters.

The presence-correction head $f_{\delta,s}$ further refines the text-initialized presence logit q_s^0 by predicting a bounded logit displacement

$$\delta_s = 2\tanh\big(f_{\delta,s}(g_s)\big) \in (-2,2)^{B\times 1}. \tag{21}$$

The refined presence logit and corresponding presence score are obtained as

$$q_s = q_s^0 + \pi_s \delta_s, \qquad p_s = \mathrm{sigm}(q_s) \in (0,1)^{B\times 1}. \tag{22}$$

Here, q_s denotes the refined target-presence logit and p_s the corresponding presence score. The reliability weight π_s regulates the image-guided correction of both the spatial parameters and the presence logit. The final layers of $f_{\Delta,s}$ and $f_{\delta,s}$ are zero-initialized, yielding $\Delta_s = 0$ and $\delta_s = 0$ at initialization. Consequently, $\Theta_s = \Theta_s^0$ and $q_s = q_s^0$ at the start of training.

Gaussian-Guided Feature Reconstruction and Update. The refined parameter vector

$$\Theta_s = [\tilde{\mu}_{x,s}, \tilde{\mu}_{y,s}, \tilde{\sigma}_{x,s}, \tilde{\sigma}_{y,s}, \tilde{\rho}_s]$$

is transformed into the bounded parameters of a correlated Gaussian. The Gaussian parameters at decoder stage s are obtained as

$$\begin{aligned} \mu_{x,s} &= \mathrm{sigm}(\tilde{\mu}_{x,s}), \qquad \mu_{y,s} = \mathrm{sigm}(\tilde{\mu}_{y,s}), \\ \sigma_{x,s} &= \sigma_{\min}^{\mathrm{dec}} + \big(\sigma_{\max}^{\mathrm{dec}} - \sigma_{\min}^{\mathrm{dec}}\big)\,\mathrm{sigm}(\tilde{\sigma}_{x,s}), \\ \sigma_{y,s} &= \sigma_{\min}^{\mathrm{dec}} + \big(\sigma_{\max}^{\mathrm{dec}} - \sigma_{\min}^{\mathrm{dec}}\big)\,\mathrm{sigm}(\tilde{\sigma}_{y,s}), \\ \rho_s &= \rho_{\max}^{\mathrm{dec}} \tanh(\tilde{\rho}_s). \end{aligned} \tag{23}$$

The Gaussian centre is $\mu_s = (\mu_{x,s}, \mu_{y,s})$, and the corresponding covariance matrix is

$$\Sigma_s = \begin{bmatrix} \sigma_{x,s}^2 & \rho_s \sigma_{x,s}\sigma_{y,s} \\ \rho_s \sigma_{x,s}\sigma_{y,s} & \sigma_{y,s}^2 \end{bmatrix}. \tag{24}$$

With $0 < \sigma_{\min}^{\mathrm{dec}} < \sigma_{\max}^{\mathrm{dec}}$ and $0 < \rho_{\max}^{\mathrm{dec}} < 1$, the transforms ensure $\sigma_{x,s}, \sigma_{y,s} > 0$ and $|\rho_s| < 1$, making Σ_s positive definite.

For a normalized spatial coordinate $t \in [0,1]^2$ on the stage grid, the Gaussian map is defined as

$$G_s(t) = \exp\left[-\frac{1}{2}(t-\mu_s)^\top \Sigma_s^{-1}(t-\mu_s)\right]. \tag{25}$$

The refined presence score modulates the Gaussian map to form the stage prior

$$P_s(t) = p_s G_s(t). \tag{26}$$

Because $0 < G_s(t) \leq 1$ and $0 < p_s < 1$, the prior satisfies $0 < P_s(t) \leq p_s < 1$ at every spatial location. Hence, P_s provides a bounded soft spatial prior that encodes the predicted target location while modulating its strength according to the target-presence score.

The stage prior is used to aggregate visual information from F_s

$$v_s = \frac{\sum_t P_s(t) F_s(t)}{\sum_t P_s(t)} \in \mathbb{R}^{B \times C_s}. \tag{27}$$

The aggregated visual representation is combined with the stage-specific text embedding and a learned stage embedding $e_s \in \mathbb{R}^{C_s}$

$$r_s = f_{r,s}\left([v_s; z_s + e_s]\right) \in \mathbb{R}^{B \times C_s}, \tag{28}$$

where $f_{r,s}$ is a two-layer MLP with GELU activation. The resulting representation is projected onto the stage grid using the same spatial prior

$$R_s(t) = P_s(t) r_s, \qquad \bar{R}_s = f_{R,s}(R_s), \tag{29}$$

where $f_{R,s}$ denotes the spatial reconstruction projection and is implemented as a 1×1 convolution followed by group normalization and GELU activation. The prior P_s therefore provides the spatial weighting for both feature aggregation and reconstruction.

A complementary text-conditioned channel residual is generated from the stage-specific text embedding

$$\begin{aligned} \gamma_s &= \tanh(z_s W_{\gamma,s}^\top + b_{\gamma,s}), \\ \beta_s &= \tanh(z_s W_{\beta,s}^\top + b_{\beta,s}), \\ U_s &= \gamma_s \odot F_s + \beta_s. \end{aligned} \tag{30}$$

where $\gamma_s, \beta_s \in \mathbb{R}^{B \times C_s}$ are broadcast over the spatial dimensions of F_s. The linear projections generating γ_s and β_s are zero-initialized, giving $U_s = 0$ at initialization.

The spatial gate for the decoder update is obtained from the current prior P_s, the upsampled preceding prior P_{2s}, and the image-guided correction weight π_s

$$A_s = \mathrm{sigm}\left(\mathrm{Conv}_{1\times1}\left([P_s; \mathrm{Up}(P_{2s}); \pi_s \mathbf{1}]\right)\right). \tag{31}$$

where $\pi_s \mathbf{1}$ denotes the correction weight broadcast over the spatial dimensions. The resulting gate satisfies $A_s \in (0,1)^{B \times 1 \times H_s \times W_s}$.

The gated decoder update is

$$\begin{aligned} \widetilde{F}_s &= F_s + \eta_s A_s \odot (\bar{R}_s + U_s), \\ D_s &= \mathrm{Refine}_s(\widetilde{F}_s). \end{aligned} \tag{32}$$

where η_s is a learnable residual strength, and $\mathrm{Refine}_s(\cdot)$ denotes a depthwise 3×3 convolution followed by a pointwise 1×1 convolution, group normalization, and GELU activation. For the intermediate decoder stages, D_s and P_s are propagated to the next finer stage, while the finest decoder feature D_4 is passed to the segmentation head.

2.4 Segmentation Head

The finest decoder feature D_4 is upsampled to the input resolution using sub-pixel convolution and mapped by a 1×1 convolution to a single-channel logit map $Z \in \mathbb{R}^{B \times 1 \times H \times W}$. The final foreground probability map is obtained as

$$\hat{Y} = \mathrm{sigm}(Z) \in (0,1)^{B \times 1 \times H \times W}. \tag{33}$$

3 Experiments and Results

3.1 Datasets and Implementation Details

Datasets. Experiments are conducted on three public medical image segmentation datasets: BKAI (Ngoc Lan et al., 2021) with 999 endoscopic polyp images, BTMRI (Cheng, 2017) with

Table 1: Quantitative comparison on BKAI, BTMRI, and BUSI datasets. Results are grouped by unimodal (×) and multimodal (✓) methods and annotated by backbone type: CNN-based (⋄), SAM-based (¶), hybrid CNN-Transformer (†), and state-space (∗∗). Each method is evaluated by trainable parameters (*M*), FLOPs (*G*), and Dice/mIoU (%). Best and second-best results are shown in **bold** and underlined, respectively.

Method	Venue	Text	Params (M) ↓	FLOPs (G) ↓	BKAI (Endoscopy)		BTMRI (MRI)		BUSI (Ultrasound)	
					Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
UNet⋄ (Ronneberger et al., 2015)	MICCAI'15	×	14.8	50.3	78.22	71.05	78.65	70.39	57.28	49.19
MT-UNet† (Wang et al., 2022a)	ICASSP'22	×	79.1	44.7	46.78	36.82	67.83	58.27	54.67	45.83
Swin-UNet† (Cao et al., 2022)	ECCV'22	×	82.3	67.3	82.68	74.78	80.64	71.98	65.18	57.96
CASCADE† (Rahman & Marculescu, 2023)	WACV'23	×	34.1	7.6	89.88	83.80	82.58	73.28	69.17	62.69
H2Former† (He et al., 2023)	IEEE TMI'23	×	33.7	24.6	77.38	69.48	80.51	71.70	63.72	56.71
MADGNet⋄ (Nam et al., 2024)	CVPR'24	×	33.4	8.8	88.81	82.50	82.77	**74.68**	68.65	62.10
Swin-UMamba∗∗ (Liu et al., 2024)	IEEE TMI'24	×	59.9	33.6	85.35	78.01	81.12	72.67	64.72	58.04
MedSA¶ (Wu et al., 2025)	MedIA'25	×	10.6	55.2	85.18	74.19	81.24	68.41	77.47	63.22
UWT-Net† (Zhang et al., 2025)	MICCAI'25	×	29.4	24.8	79.50	71.81	80.21	71.54	63.51	56.46
EccoMamba∗∗ (Xu et al., 2026)	AAAI'26	×	38.3	**6.2**	86.83	79.64	74.07	65.15	65.38	58.90
CRIS⋄ (Wang et al., 2022b)	CVPR'22	✓	161.0	21.1	81.56	68.86	70.23	55.67	78.05	65.26
TGANet⋄ (Tomar et al., 2022)	MICCAI'22	✓	19.8	41.9	86.60	80.89	81.50	72.91	69.33	62.32
BKINet† (Ding et al., 2023)	IEEE TMM'23	✓	156.4	83.1	85.67	74.94	80.10	66.81	83.15	71.16
C2FVL† (Shan et al., 2023)	ICASSP'23	✓	24.0	29.7	85.81	75.15	79.79	66.38	79.33	65.74
LViT† (Li et al., 2023)	IEEE TMI'23	✓	29.7	54.1	85.59	74.81	81.23	68.40	80.86	67.87
MMI-UNet† (Bui et al., 2024)	MICCAI'24	✓	56.2	22.1	90.24	82.22	82.21	69.80	78.97	65.25
RecLMIS† (Huang et al., 2024)	IEEE TMI'24	✓	23.7	24.1	81.44	68.69	79.94	66.59	82.86	70.74
ReMamber∗∗ (Yang et al., 2024)	ECCV'24	✓	102.5	7.4	83.32	71.41	79.30	65.70	84.87	73.71
DETRIS† (Huang et al., 2025)	AAAI'25	✓	25.5	33.0	89.53	81.05	82.30	69.93	82.23	69.83
ProLearn† (Ye et al., 2025)	ICCV'25	✓	61.2	76.3	90.90	83.49	82.36	70.26	82.04	69.74
TeViA† (Zeng et al., 2025)	IEEE TMI'25	✓	146.8	11.2	90.49	82.64	83.95	72.34	81.02	68.10
MedCLIPSeg† (Koleilat et al., 2026)	CVPR'26	✓	18.7	73.6	84.28	72.84	77.19	62.86	84.04	72.47
UA-VLS∗∗ (Das et al., 2026)	WACV'26	✓	39.9	17.8	87.52	77.82	81.74	69.12	84.37	72.97
ViLMedSAM¶ (Qian et al., 2026)	CVPR'26	✓	**3.9**	394.5	67.13	50.53	67.24	50.65	74.66	59.57
TIGR-Seg†	(Ours)	✓	4.6	18.5	**92.78 ± 0.38**	**86.53 ± 0.57**	**84.76 ± 0.39**	73.56 ± 0.55	**86.77 ± 0.44**	**76.63 ± 0.66**

3,064 brain MRI images, and BUSI (Al-Dhabyani et al., 2020) with 780 breast ultrasound images. Each dataset uses a fixed 80/20 train-test split. Images are paired with structured textual descriptions containing target class, coarse location, count, shape, and appearance. Text annotations for BTMRI are obtained from (Koleilat et al., 2026), while BKAI and BUSI use annotations from (Poudel et al., 2024). Cross-dataset evaluation additionally considers Kvasir-SEG (Jha et al., 2019) and BUS-BRA (Gómez-Flores et al., 2024), with textual descriptions obtained from (Poudel et al., 2024) and (Koleilat et al., 2026), respectively.

Implementation Details. All experiments were implemented in PyTorch on an NVIDIA A100 SXM4 GPU with 40 GB memory. Images and masks were resized to 224×224, with random zoom applied with probability 0.1. GGFA used a bottleneck dimension of $r = 4$. AdamW was used with batch size 32 and cosine-annealed learning rate from 3×10^{-4} to 1×10^{-6}. Training used equally weighted binary cross-entropy and soft Dice losses. From each training partition, 10% was reserved for validation. Models were trained for up to 200 epochs with early stopping after 30 epochs without improvement in validation Dice; the checkpoint with the highest validation Dice was used for test evaluation. Results are reported as mean ± standard deviation over three seeds $\{25, 250, 2500\}$. Following prior text-guided segmentation studies (Li et al., 2023; Huang et al., 2024), performance was evaluated using Dice and mean Intersection over Union (mIoU).

3.2 Comparison with State-of-the-Art Methods

Table 1 compares TIGR-Seg with both vision-only and vision-language segmentation methods spanning CNN-based, SAM-based, hybrid CNN-Transformer, and state-space architectures. All compared methods were reimplemented using the same fixed train-test partitions, while method-specific architectures, pretrained initialization, optimization settings, and training procedures followed the corresponding original papers. The same Dice and mIoU computation procedure was used for evaluation. TIGR-Seg achieves the highest Dice on BKAI, BTMRI, and BUSI, and the highest mIoU on BKAI and BUSI among the compared methods, with 4.6M trainable parameters and 18.5G FLOPs. Additional qualitative comparisons are provided in Supplementary Sec. C.

Table 2: **Ablation of GGFA and the TIGR Decoder on BKAI and BUSI.** All variants use the same base architecture, with ✓ indicating inclusion of the corresponding component. Dice, mIoU, and HD95 are reported.

GGFA	TIGR	BKAI			BUSI		
		Dice(%) ↑	mIoU(%) ↑	HD95 ↓	Dice(%) ↑	mIoU(%) ↑	HD95 ↓
✓	×	89.00	82.92	30.27	83.11	71.10	21.60
×	✓	91.16	83.77	19.42	84.62	73.34	15.98
✓	✓	**92.78**	**86.53**	**18.43**	**86.77**	**76.63**	**14.82**

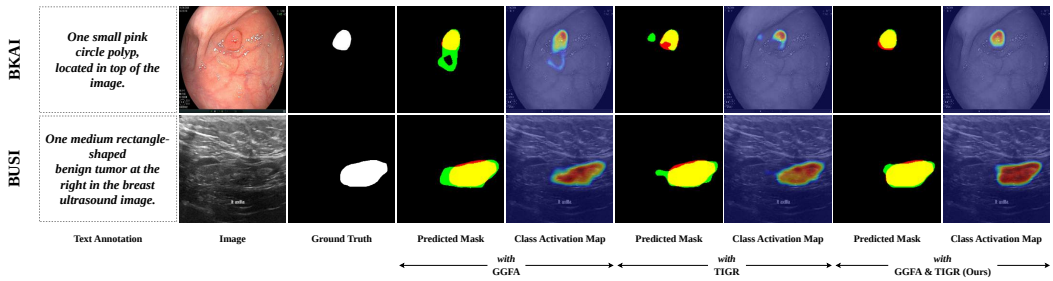


Figure 3: **Qualitative ablation of GGFA and the TIGR Decoder on BKAI and BUSI.** Segmentation overlays and activation maps are shown for GGFA alone, the TIGR Decoder alone, and their combination. Yellow, red, and green denote true-positive, false-negative, and false-positive regions, respectively.

4 Ablation Studies

The effectiveness of the proposed components and design choices of TIGR-Seg is evaluated through ablation analysis. Additional studies are provided in Supplementary Sec. D.

4.1 Effectiveness of GGFA and TIGR Decoder

Table 2 evaluates the individual and combined effects of GGFA and the TIGR Decoder on BKAI and BUSI. With GGFA enabled but without the TIGR Decoder, Dice scores of 89.00% on BKAI and 83.11% on BUSI are obtained, indicating that GGFA alone provides effective adaptation of the frozen vision backbone. When GGFA is removed and the Swin encoder remains fully frozen, the TIGR Decoder achieves 91.16% and 84.62% Dice, corresponding to gains of 2.16 and 1.51 percentage points over the GGFA-only variant. Combining both components further improves Dice to 92.78% and 86.77%, yielding additional gains of 1.62 and 2.15 percentage points over the TIGR-only variant. The consistent improvements in Dice, mIoU, and HD95 indicate complementary contributions from spatially aware encoder adaptation and text-guided decoder refinement. Fig. 3 further shows improved target delineation with fewer false-negative and false-positive regions when both components are used.

5 Conclusion

This work presents TIGR-Seg, a parameter-efficient text-guided framework for medical image segmentation. GGFA adapts a frozen vision encoder through context-conditioned feature modulation and dynamic Gaussian spatial gating, while the TIGR Decoder initializes spatial guidance from text and progressively refines it using visual evidence. The resulting spatial priors guide feature aggregation and multi-scale decoder refinement. Experiments across endoscopy, brain MRI, and breast ultrasound demonstrate strong segmentation performance with only 4.6M trainable parameters. The presented studies support the complementary contributions of GGFA and the TIGR Decoder, while reduced-data and cross-dataset evaluations further demonstrate the effectiveness of the framework under limited training data and across related target datasets. Overall, TIGR-Seg provides an effective approach for combining parameter-efficient visual adaptation with image-refined textual guidance for medical image segmentation.

AI USE STATEMENT

In this work, generative AI tools were not used for tasks requiring disclosure under the ICLR 2027 AI Policy. Generative AI tools were used only for language editing of the manuscript. All AI-assisted text was reviewed and revised by the authors for technical accuracy, clarity, and consistency with the conducted research. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses publicly available medical image segmentation datasets and does not involve new data collection or direct interaction with human participants. All datasets and associated resources are cited to their original sources. The experiments and reported results were conducted and presented in accordance with standard research-integrity practices.

REPRODUCIBILITY STATEMENT

The experiments in this work are fully reproducible. All datasets and code used for model training and evaluation is publicly available at: `https://anonymous.4open.science/r/tigr-seg-E3D8`. Detailed information on the experimental setup, model configurations, and hyperparameters is provided to enable independent reproduction of results.

REFERENCES

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

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 46:10076–10095, 2024a.

Reza Azad, Amirhossein Kazerouni, Moein Heidari, Ehsan Khodapanah Aghdam, Amirali Molaei, Yiwei Jia, Abin Jose, Rijo Roy, and Dorit Merhof. Advances in medical image analysis with vision transformers: a comprehensive review. *Medical Image Analysis (MedIA)*, 91:103000, 2024b.

Phuoc-Nguyen Bui, Duc-Tai Le, and Hyunseung Choo. Visual-textual matching attention for lesion segmentation in chest images. In *27th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 702–711. Springer, 2024.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *17th European Conference on Computer Vision (ECCV)*, pp. 205–218. Springer, 2022.

Jun Cheng. Brain tumor dataset, 2017. URL `https://doi.org/10.6084/m9.figshare.1512427.v5`.

Pierre-Henri Conze, Gustavo Andrade-Miranda, Vivek Kumar Singh, Vincent Jaouen, and Dimitris Visvikis. Current and emerging trends in medical image segmentation with deep learning. *IEEE Transactions on Radiation and Plasma Medical Sciences (IEEE TRPMS)*, 7:545–569, 2023.

Aryan Das, Tanishq Rachamalla, Koushik Biswas, Swalpa Kumar Roy, and Vinay Kumar Verma. Uncertainty-aware vision-language segmentation for medical imaging. In *18th IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8490–8499, 2026.

Manish Dhakal, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Vlsm-adapter: Finetuning vision-language segmentation efficiently with lightweight blocks. In *27th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 712–722. Springer, 2024.

Haixin Ding, Shengchuan Zhang, Qiong Wu, Songlin Yu, Jie Hu, Liujuan Cao, and Rongrong Ji. Bilateral knowledge interaction network for referring image segmentation. *IEEE Transactions on Multimedia (IEEE TMM)*, 26:2966–2977, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations (ICLR)*, 2021.

Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In *7th International Conference on Medical Imaging with Deep Learning (MIDL)*, pp. 406–425. PMLR, 2024.

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*, 51:3110–3123, 2024.

Jie Gui, Hang Tu, Wen Sha, and Xiuquan Du. Towards text-mask consistency in medical image segmentation. In *14th International Conference on Learning Representations (ICLR)*, 2026.

Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. *IEEE Transactions on Medical Imaging (IEEE TMI)*, 42(9):2763–2775, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *36th International Conference on Machine Learning (ICML)*, pp. 2790–2799. PMLR, 2019.

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 *10th International Conference on Learning Representations (ICLR)*, 2022.

Jiaqi Huang, Zunnan Xu, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Densely connected parameter-efficient tuning for referring image segmentation. In *39th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3653–3661, 2025.

Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging (IEEE TMI)*, 44:1821–1835, 2024.

Soojin Hwang, Jaeyoon Sim, and Won Hwa Kim. Himix: Hierarchical visual-textual mixing network for lesion segmentation. In *18th IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5332–5341, 2026.

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 *25th International Conference on MultiMedia Modeling (MMM)*, pp. 451–462. Springer, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *19th IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

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 *43rd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1406–1417, 2026.

David B Larson, Alex J Towbin, Rebecca M Pryor, and Lane F Donnelly. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. *Radiology*, 267:240–250, 2013.

Xiang Li, Like Li, Yuchen Jiang, Hao Wang, Xinyu Qiao, Ting Feng, Hao Luo, and Yong Zhao. Vision-language models in medical image analysis: From simple fusion to general large models. *Information Fusion*, 118:102995, 2025.

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 (IEEE TMI)*, 43:96–107, 2023.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis (MedIA)*, 42:60–88, 2017.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Lequan Yu, Yong Liang, Yizhou Yu, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation. *IEEE Transactions on Medical Imaging (IEEE TMI)*, 44:3898–3908, 2024.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. A generalist medical language model for disease diagnosis assistance. *Nature Medicine*, pp. 1–11, 2025.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *39th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, 2022.

Yinbin Lu and Alan Wang. Integrating language into medical visual recognition and reasoning: A survey. *Medical Image Analysis (MedIA)*, 102:103514, 2025.

Ju-Hyeon Nam, Nur Suriza Syazwany, Su Jung Kim, and Sang-Chul Lee. Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In *41st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11480–11491, 2024.

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 *16th International Symposium on Visual Computing (ISVC)*, pp. 15–28. Springer, 2021.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *32nd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3942–3951, 2018.

Kanchan Poudel, Manish Dhakal, Prasiddha Bhandari, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Exploring transfer learning in medical image segmentation using vision-language models. In *7th International Conference on Medical Imaging with Deep Learning (MIDL)*, 2024.

Abdul Qayoom, Juanying Xie, and Haider Ali. Polyp segmentation in medical imaging: challenges, approaches and future directions. *Artificial Intelligence Review*, 58:169, 2025.

Chengcan Qian, Dong Nie, Geng Chen, Daoqiang Zhang, and Xuyun Wen. Simple-vilmedsam: Simple text prompts meet vision-language models for medical image segmentation. In *43rd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30042–30052, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *38th International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *15th IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6222–6231, 2023.

Praveenbalaji Rajendran, Mojtaba Safari, Wenfeng He, Mingzhe Hu, Keyur Shah, Shansong Wang, Jun Zhou, and Xiaofeng Yang. Foundation models in medical image analysis: A systematic review and quantitative analysis. *Medical Image Analysis (MedIA)*, 115:104293, 2026.

Ramin Ranjbarzadeh, Annalina Caputo, Erfan Babaee Tirkolaee, Saeid Jafarzadeh Ghoushchi, and Malika Bendechache. Brain tumor segmentation of mri images: A comprehensive review on the application of artificial intelligence tools. *Computers in Biology and Medicine*, 152:106405, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *13th International Conference on Learning Representations (ICLR)*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015.

Dandan Shan, Zihan Li, Wentao Chen, Qingde Li, Jie Tian, and Qingqi Hong. Coarse-to-fine covid-19 segmentation via vision-language alignment. In *48th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: Text-guided attention for improved polyp segmentation. In *25th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 151–160. Springer, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *38th International Conference on Machine Learning (ICML)*, pp. 10347–10357. PMLR, 2021.

Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2390–2394. IEEE, 2022a.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *39th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11686–11695, 2022b.

Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis (MedIA)*, 102:103547, 2025.

Qingling Xia, Hong Zheng, Haonan Zou, Dinghao Luo, Hongan Tang, Lingxiao Li, and Bin Jiang. A comprehensive review of deep learning for medical image segmentation. *Neurocomputing*, 613: 128740, 2025.

Min Xian, Yingtao Zhang, Heng-Da Cheng, Fei Xu, Boyu Zhang, and Jianrui Ding. Automatic breast ultrasound image segmentation: A survey. *Pattern Recognition*, 79:340–355, 2018.

Junlin Xu, Jincan Li, Feifei Cui, Zhuang Zhang, Jialiang Yang, Shuting Jin, Qiangguo Jin, and Yajie Meng. Eccomamba: Enhanced cross-hierarchical continuity orthogonal mamba for medical image segmentation. In *40th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 11287–11295, 2026.

Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. In *18th European Conference on Computer Vision (ECCV)*, pp. 108–126. Springer, 2024.

Shuchang Ye, Usman Naseem, Mingyuan Meng, and Jinman Kim. Alleviating textual reliance in medical language-guided segmentation via prototype-driven semantic approximation. In *20th IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22316–22326, 2025.

Qingjie Zeng, Huan Luo, Zilin Lu, Yutong Xie, Zhiyong Wang, Yanning Zhang, and Yong Xia. Harnessing text insights with visual alignment for medical image segmentation. *IEEE Transactions on Medical Imaging (IEEE TMI)*, 45:477–489, 2025.

Pengcheng Zhang, Xiaocao Ouyang, and Ran Peng. Uwt-net: Mining low-frequency feature information for medical image segmentation. In *28th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 615–624. Springer, 2025.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22: 166–176, 2025.

TIGR-SEG: TEXT-INITIALIZED GAUSSIAN REFINEMENT WITH SPATIALLY AWARE VISUAL ADAPTATION FOR MEDICAL IMAGE SEGMENTATION

SUPPLEMENTARY MATERIAL

A EXTENDED RELATED WORK

A.1 VISION-ONLY SEGMENTATION

Vision-only medical image segmentation has evolved from convolutional encoder-decoder architectures to Transformer-based, Mamba-based, and foundation-model-based designs (Xia et al., 2025), (Rajendran et al., 2026). U-Net (Ronneberger et al., 2015) established the widely adopted U-shaped encoder-decoder architecture with skip connections for combining features across multiple resolutions. Transformer-based approaches such as MT-UNet (Wang et al., 2022a), Swin-UNet (Cao et al., 2022), CASCADE (Rahman & Marculescu, 2023), and H2Former (He et al., 2023) further improve global context modeling, hierarchical feature representation, and multi-scale feature refinement. Frequency-aware methods, including MADGNet (Nam et al., 2024) and UWT-Net (Zhang et al., 2025), incorporate information from different spatial-frequency components to enhance structural representation. More recently, Swin-UMamba (Liu et al., 2024) and EccoMamba (Xu et al., 2026) employ Mamba-based architectures to capture long-range dependencies efficiently. PEFT has also been explored for adapting pretrained foundation models while preserving most backbone parameters. Med-SA (Wu et al., 2025) adapts SAM (Kirillov et al., 2023) through bottleneck adapters and a Hyper-Prompting Adapter (HyP-Adpt) for prompt-conditioned adaptation. While PEFT reduces the number of trainable parameters required for adaptation, dense medical image segmentation can further benefit from spatially aware adaptation mechanisms that account for both feature information and spatial structure.

A.2 VISION-LANGUAGE SEGMENTATION

Recent vision-language segmentation methods have explored different strategies for incorporating textual information into medical image segmentation (Li et al., 2025). LViT (Li et al., 2023) combines convolutional and Transformer representations with textual guidance, while RecLMIS (Huang et al., 2024) improves image-text consistency through cross-modal conditioned reconstruction, C2Seg (Gui et al., 2026) enhances text–mask consistency through cluster-aware contrastive learning and bidirectional complementary attention, and TeViA (Zeng et al., 2025) introduces segmentation-specific text-to-vision alignment. Mamba-based multimodal methods have also been explored, with UA-VLS (Das et al., 2026) using a lightweight State Space Mixer for efficient cross-modal fusion and uncertainty-aware segmentation, and ReMamber (Yang et al., 2024) introducing a Mamba Twister for long-range visual-textual interaction. CLIP-based approaches leverage pretrained image-text representations for segmentation; MedCLIPSeg (Koleilat et al., 2026) uses probabilistic patch-level image-text interaction with uncertainty modeling, while Simple-ViLMedSAM (Qian et al., 2026) integrates CLIP (Radford et al., 2021) and SAM (Kirillov et al., 2023) for segmentation from simple textual queries. Parameter-efficient adaptation has also been explored, with DETRIS (Huang et al., 2025) combining densely connected low-rank visual adaptation with text adapters, while VLSM-Adapter (Dhakal et al., 2024) inserts bottleneck adapters into intermediate visual and textual representations of pretrained CLIP-based segmentation models. Despite these advances, textual information is commonly incorporated through feature interaction, alignment, or conditioning, without explicitly modeling its coarse spatial guidance as a hypothesis that can be refined using image evidence. This motivates a formulation in which text provides an initial spatial hypothesis that is subsequently verified and refined using visual evidence before guiding segmentation.

Algorithm 1 Gaussian-Gated Feature Adaptation (**GGFA**)

Input: $X \in \mathbb{R}^{B_w \times N_w \times d_{\text{in}}}$, $N_w = M^2$
Frozen base: $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, $b_0 \in \mathbb{R}^{d_{\text{out}}}$
Token coordinate: $t_i \in [0,1]^2$, $i = 1, \dots, N_w$
Bounds: $0 < \sigma_{\min} < \sigma_{\max}$, $0 < \rho_{\max} < 1$
B_w : number of windows, r : context dimension

class GGFA(*Module*):
  def __init__(self, d_{in}, d_{out}, r, W_0, b_0):
    super().__init__()
    self.$W_0 \leftarrow W_0$ — $[d_{\text{out}}, d_{\text{in}}]$, frozen
    self.$b_0 \leftarrow b_0$ — $[d_{\text{out}}]$, frozen
    # trainable shared context and modulation
    self.$W_d \leftarrow \text{Parameter}(d_{\text{in}}, r)$ — $[d_{\text{in}}, r]$
    self.LN $\leftarrow \text{LayerNorm}(r)$
    self.$f_\gamma \leftarrow \text{Linear}(r, d_{\text{out}})$ — zero-initialized
    self.$f_\beta \leftarrow \text{Linear}(r, d_{\text{out}})$ — zero-initialized
    # trainable Gaussian-gating branch
    self.$f_G \leftarrow [\text{Linear}(r, r), \text{GELU}, \text{Linear}(r, 6)]$
    self.$\lambda \leftarrow \lambda_{\text{init}}$ — trainable scalar
  def forward(self, X):
    # 1. frozen pretrained response
    $Y_0 \leftarrow X W_0^\top + b_0$ — $[B_w, N_w, d_{\text{out}}]$
    # 2. shared window context
    $h \leftarrow \text{GELU}\Big(\text{LN}\Big(\frac{1}{N_w}\sum_{i=1}^{N_w} X_i W_d\Big)\Big)$ — $[B_w, r]$
    # 3. context-conditioned feature modulation
    $\gamma \leftarrow \tanh(f_\gamma(h)), \quad \beta \leftarrow \tanh(f_\beta(h))$ — $[B_w, d_{\text{out}}]$
    $Y_{\text{mod}} \leftarrow \gamma \odot Y_0 + \beta$ — $[B_w, N_w, d_{\text{out}}]$
    # 4. dynamic Gaussian gate
    $[\tilde{\mu}_x, \tilde{\mu}_y, \tilde{\sigma}_x, \tilde{\sigma}_y, \tilde{\rho}, \tilde{a}] \leftarrow f_G(h)$
    $\mu_x \leftarrow \text{sigm}(\tilde{\mu}_x), \quad \mu_y \leftarrow \text{sigm}(\tilde{\mu}_y)$
    $\sigma_x \leftarrow \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\text{sigm}(\tilde{\sigma}_x)$
    $\sigma_y \leftarrow \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\text{sigm}(\tilde{\sigma}_y)$
    $\rho \leftarrow \rho_{\max} \tanh(\tilde{\rho}), \quad a \leftarrow \text{sigm}(\tilde{a})$
    $\mu \leftarrow (\mu_x, \mu_y), \quad \Sigma \leftarrow \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$
    for $i = 1, \dots, N_w$:
      $G_i \leftarrow a \exp\big[-\frac{1}{2}(t_i - \mu)^\top \Sigma^{-1}(t_i - \mu)\big]$
    $G \leftarrow \{G_i\}_{i=1}^{N_w}$ — $[B_w, N_w, 1]$
    # 5. Gaussian-gated residual adaptation
    return $Y_0 + \lambda\, G \odot Y_{\text{mod}}$ — $[B_w, N_w, d_{\text{out}}]$

B GGFA Algorithm and Parameter Complexity

Algorithm 1 provides the PyTorch-style pseudocode for GGFA.

For one adapted projection, the context projection and LayerNorm contribute $d_{\text{in}}r + 2r$ trainable parameters, the scale and shift heads contribute $2(rd_{\text{out}} + d_{\text{out}})$, the two-layer Gaussian predictor contributes $(r^2 + r) + (6r + 6)$, and the learnable residual strength λ contributes one parameter. Thus,

$$P_{\text{GGFA}} = d_{\text{in}}r + 2r\, d_{\text{out}} + r^2 + 9r + 2d_{\text{out}} + 7. \tag{34}$$

Full fine-tuning of the same projection requires

$$P_{\text{full}} = d_{\text{in}}d_{\text{out}} + d_{\text{out}}. \tag{35}$$

For $d_{\text{in}} = d_{\text{out}} = d$ and $r \ll d$, GGFA therefore scales as $\mathcal{O}(rd)$ in trainable parameters, compared with $\mathcal{O}(d^2)$ for full fine-tuning.

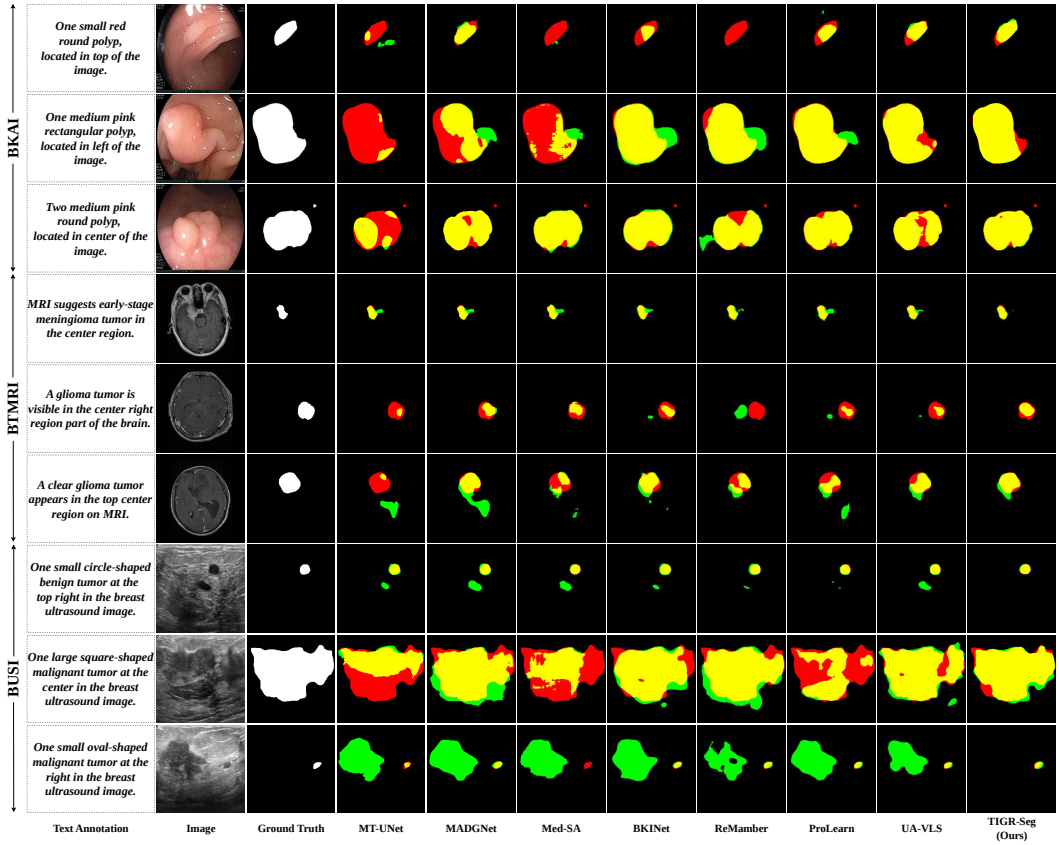


Figure 4: **Qualitative segmentation comparison on BKAI, BTMRI, and BUSI.** The overlays show true-positive regions in yellow, false-negative regions in red, and false-positive regions in green.

C Qualitative Segmentation Results

Fig. 4 provides qualitative comparisons with MT-UNet (Wang et al., 2022a), MADGNet (Nam et al., 2024), Med-SA (Wu et al., 2025), BKINet (Ding et al., 2023), ReMamber (Yang et al., 2024), ProLearn (Ye et al., 2025), and UA-VLS (Das et al., 2026) across BKAI, BTMRI, and BUSI. The predictions illustrate differences in target localization and boundary delineation across the compared methods, with TIGR-Seg showing fewer false-negative and false-positive regions in the presented examples.

D Additional Ablation Studies

D.1 Analysis of Progressive Gaussian Refinement

Figure 5 illustrates the evolution of the refined Gaussian spatial prior across the TIGR decoder stages at output strides $s \in \{16, 8, 4\}$. The coarse-stage prior provides an initial spatial localization, while subsequent stages progressively adjust its center, spatial extent, and orientation using finer visual evidence. The resulting priors become more closely aligned with the target regions in both BKAI and BUSI examples, demonstrating how the text-initialized spatial hypothesis is progressively refined rather than used as fixed localization guidance. The corresponding segmentation results further show how this stage-wise refinement contributes to accurate target delineation.

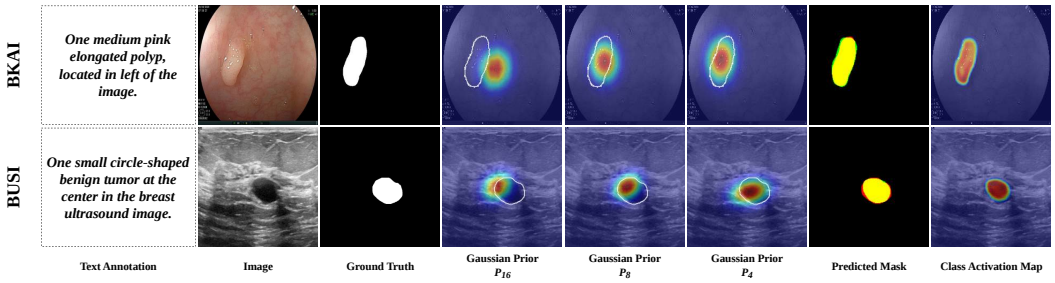


Figure 5: **Progressive Gaussian refinement across TIGR Decoder stages.** Refined Gaussian spatial priors are visualized across successive decoder stages for representative BKAI and BUSI cases, with the ground-truth boundary indicated in white. The final segmentation overlay shows true-positive, false-negative, and false-positive regions in yellow, red, and green, respectively, followed by the corresponding class activation map.

Table 3: **Effect of TIGR Decoder components on BKAI and BUSI.** Components are progressively added to evaluate their contributions using Dice and mIoU.

Text-Conditioned Feature Modulation	Text-Guided Gaussian Proposal	Image-Guided Gaussian Correction	BKAI		BUSI	
			Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
✓	×	×	89.04	80.25	83.13	71.13
✓	✓	×	91.35	84.10	84.75	73.54
✓	✓	✓	**92.78**	**86.53**	**86.77**	**76.63**

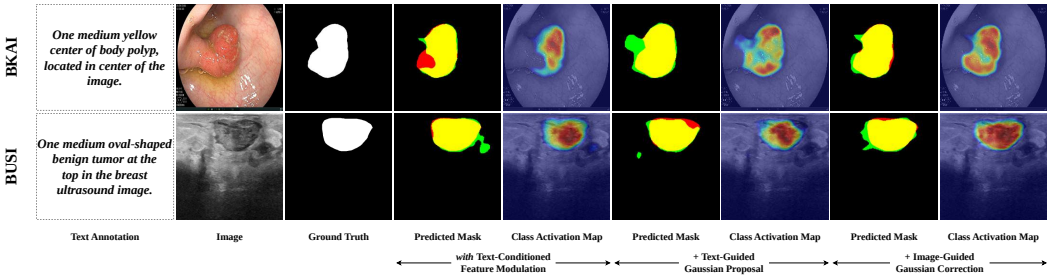


Figure 6: **Qualitative ablation of TIGR Decoder components on BKAI and BUSI.** Results are shown for progressively adding Text-Conditioned Feature Modulation, Text-Guided Gaussian Proposal, and Image-Guided Gaussian Correction. Yellow, red, and green denote true-positive, false-negative, and false-positive regions, respectively, together with the corresponding class activation maps.

D.2 Analysis of TIGR Decoder Components

Table 3 evaluates the contribution of the main components of the TIGR Decoder on BKAI and BUSI. Text-conditioned feature modulation provides the initial performance, which improves further with the addition of the Text-Guided Gaussian Proposal. Incorporating Image-Guided Gaussian Correction yields the best Dice and mIoU on both datasets. These results indicate complementary contributions from text-guided spatial initialization and image-guided refinement. Fig. 6 provides the corresponding qualitative comparison across the component configurations.

D.3 Analysis of GGFA Components

Table 4 analyzes the contribution of the main components of GGFA on BKAI and BUSI. Starting from the fully frozen vision encoder, feature modulation is first introduced to provide context-conditioned feature adaptation. Gaussian gating is then incorporated to spatially control the feature correction, with the correlation term disabled by setting $\rho = 0$. Finally, enabling the learned correlation allows the Gaussian gate to model dependencies between the two spatial axes. Performance

Table 4: **Effect of GGFA components on BKAI and BUSI.** Components are progressively incorporated and evaluated using Dice and mIoU.

Feature Modulation	Gaussian Gating	Gaussian Correlation	BKAI		BUSI	
			Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
×	×	×	91.16	83.77	84.62	73.34
✓	×	×	91.63	84.55	85.18	74.19
✓	✓	×	92.31	85.72	86.02	75.47
✓	✓	✓	**92.78**	**86.53**	**86.77**	**76.63**

Table 5: **Effect of text prompt variations on BKAI and BUSI.** Results are reported using Dice and mIoU.

Text Prompt Design	BKAI		BUSI	
	Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
Contradictory	76.34	62.11	65.92	49.18
No Text	79.48	66.97	71.26	55.43
Underdescriptive	84.13	73.31	77.54	63.42
Overdescriptive	89.67	81.27	82.91	70.81
Original	**92.78**	**86.53**	**86.77**	**76.63**

improves progressively as these components are incorporated, with the complete GGFA configuration achieving the highest Dice and mIoU on both datasets. These results indicate complementary contributions from feature modulation, spatial gating, and Gaussian correlation.

D.4 Effect of Text Prompt Variations

Table 5 evaluates the effect of text availability and prompt formulation on TIGR-Seg using BKAI and BUSI. The Original prompt contains both target description and spatial information, e.g., *"One small, circle-shaped benign tumor is located at the top right of the breast ultrasound image."* In the No Text setting, textual guidance is removed entirely. The Underdescriptive setting retains only a simplified target description while omitting detailed shape and location information, e.g., *"A small benign tumor is visible in the breast ultrasound image."* The Overdescriptive setting preserves the same correct semantic and spatial information but expresses it with additional detail, e.g., *"A single, relatively small benign tumor with a circular shape is visible in the breast ultrasound image and is situated within the upper-right region of the image."* The Contradictory setting retains the target attributes but provides an incorrect spatial location, e.g., *"One small, circle-shaped benign tumor is located at the bottom left of the breast ultrasound image."* Underdescriptive and overdescriptive prompts remain effective, while the Original prompt achieves the highest performance on both datasets, indicating robustness to variations in prompt detail. Removing textual guidance further reduces performance, whereas contradictory text produces the largest degradation. This pattern suggests that TIGR-Seg actively uses the semantic and spatial information provided by the text rather than treating it as incidental input.

D.5 Comparison with PEFT Methods

Table 6 compares GGFA with Houlsby Adapter (Houlsby et al., 2019), LoRA (Hu et al., 2022), FiLM (Perez et al., 2018), and full fine-tuning on BKAI and BUSI. For this comparison, GGFA is replaced by the corresponding adaptation strategy while the remaining model configuration is kept unchanged. GGFA achieves the highest Dice and mIoU on both datasets with 4.6M trainable parameters. Compared with full fine-tuning, GGFA improves Dice by 3.69 percentage points on BKAI and 0.97 percentage points on BUSI while reducing the number of trainable parameters from 32.0M to 4.6M. The qualitative results in Fig. 7 provide the corresponding comparison of the predicted segmentation regions across the different adaptation strategies.

Table 6: **Comparison of adaptation strategies on BKAI and BUSI.** Results are reported using trainable parameters, Dice, and mIoU.

PEFT Type	Params (M) ↓	BKAI		BUSI	
		Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
Houlsby Adapter (Houlsby et al., 2019)	4.7	88.72	79.73	84.77	73.57
LoRA (Hu et al., 2022)	4.5	87.76	78.20	85.49	74.66
FiLM (Perez et al., 2018)	4.6	87.50	77.77	84.88	73.73
Full-finetune	32.0	89.09	80.33	85.80	75.13
GGFA (Ours)	**4.6**	**92.78**	**86.53**	**86.77**	**76.63**

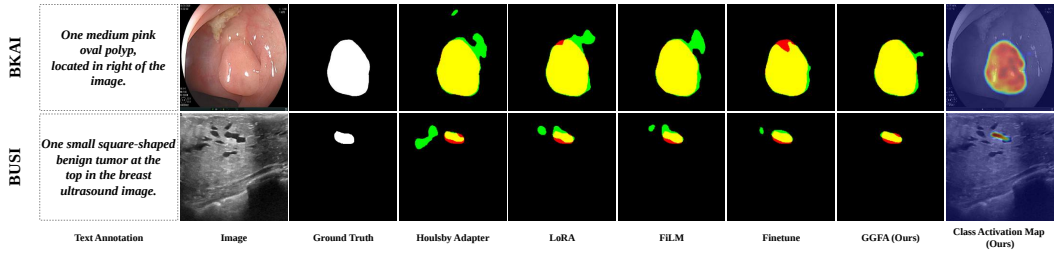


Figure 7: **Qualitative comparison of adaptation strategies on BKAI and BUSI.** Segmentation results are shown for Houlsby Adapter, LoRA, FiLM, full fine-tuning, and GGFA. Yellow, red, and green denote true-positive, false-negative, and false-positive regions, respectively. The final column shows the corresponding class activation map for the GGFA-based model.

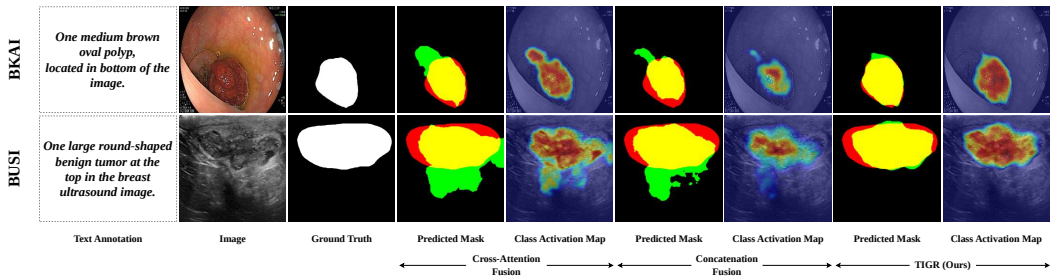


Figure 8: **Qualitative comparison of decoder designs.** Results are shown for BKAI and BUSI cases using Cross-Attention Fusion, Concatenation Fusion, and the TIGR Decoder. Yellow, red, and green denote true-positive, false-negative, and false-positive regions, respectively, together with the corresponding class activation maps.

D.6 Effect of Decoder Design

Table 7 compares the TIGR Decoder with Concatenation Fusion and Cross-Attention Fusion on BKAI and BUSI. The TIGR Decoder achieves the highest Dice and mIoU and the lowest HD95 across both datasets. Compared with Cross-Attention Fusion, the TIGR Decoder improves Dice by 5.51 and 2.99 percentage points on BKAI and BUSI, respectively, while reducing HD95 from 25.08 to 18.43 on BKAI and from 22.65 to 14.82 on BUSI. These results indicate that the proposed decoder design provides consistent improvements over the two alternative fusion strategies considered in this ablation. Fig. 8 provides the corresponding qualitative comparison through segmentation overlays and class activation maps.

D.7 Cross-Dataset Generalization

Table 8 evaluates cross-dataset generalization from BKAI to Kvasir-SEG and from BUSI to BUS-BRA. TIGR-Seg achieves the highest Dice and mIoU in both settings. Compared with ProLearn, the strongest competing method, TIGR-Seg improves Dice by 2.88 percentage points and mIoU by 3.55 percentage points on Kvasir-SEG, and by 4.17 and 6.04 percentage points, respectively, on

Table 7: **Comparison of decoder designs on BKAI and BUSI.** Results are reported using Dice, mIoU, and HD95.

Decoder	BKAI			BUSI		
	Dice(%) ↑	mIoU(%) ↑	HD95 ↓	Dice(%) ↑	mIoU(%) ↑	HD95 ↓
Concatenation Fusion	86.08	75.78	27.02	82.69	70.49	27.40
Cross-Attention Fusion	87.27	77.46	25.08	83.78	72.09	22.65
TIGR (Ours)	**92.78**	**86.53**	**18.43**	**86.77**	**76.63**	**14.82**

Table 8: **Cross-dataset generalization on Kvasir-SEG and BUS-BRA.** Models are evaluated on the target datasets without fine-tuning.

Method	BKAI → Kvasir-SEG		BUSI → BUS-BRA	
	Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
CRIS (Wang et al., 2022b)	78.97	65.34	63.45	48.20
LViT (Li et al., 2023)	65.61	48.82	61.00	43.88
ReMamber (Yang et al., 2024)	74.25	59.05	61.39	44.29
ProLearn (Ye et al., 2025)	79.82	66.96	80.38	67.20
MedCLIPSeg (Koleilat et al., 2026)	78.05	64.01	73.84	58.53
TIGR-Seg (Ours)	**82.70**	**70.51**	**84.55**	**73.24**

BUS-BRA. These results indicate improved generalization across related datasets with different data distributions.

D.8 Effect of Training Data Size

Table 9 evaluates the effect of training data size on BKAI and BUSI. All methods are evaluated on the same test set. LViT and RecLMIS are trained using the complete training set, while TIGR-Seg is evaluated using 10%, 25%, 50%, and 100% of the available training data. Performance consistently improves as the amount of training data increases. Notably, with only 50% of the training data, TIGR-Seg achieves higher Dice and mIoU than both LViT and RecLMIS trained on the full training set. With 50% of the training data, TIGR-Seg improves Dice by 1.66 percentage points over LViT on BKAI and by 0.83 percentage points over RecLMIS on BUSI. These results indicate favorable performance under reduced training data.

D.9 Analysis of Image Encoders

Table 10 compares different frozen image encoders on BKAI and BUSI, while keeping GGFA and the TIGR Decoder unchanged. The Swin-based configuration records the highest Dice and mIoU among the evaluated encoders on both datasets, while ViT, DeiT, and SAM2 yield lower performance. These results support the effectiveness of using Swin with GGFA and the TIGR Decoder.

D.10 Model Efficiency Analysis

Table 11 compares the computational efficiency of TIGR-Seg with existing segmentation methods. TIGR-Seg requires only 4.6M trainable parameters, the lowest among the compared methods, despite using frozen vision and text encoders with 166.9M total parameters. Its computational cost is 18.5 GFLOPs, which remains substantially lower than Med-SA, ProLearn, and MedCLIPSeg, although several lighter vision-only and state-space models require fewer FLOPs. TIGR-Seg also shows lower inference latency than ProLearn and MedCLIPSeg, while requiring higher latency than the remaining compared methods. Overall, the results show that TIGR-Seg maintains a low trainable-parameter requirement with moderate computational cost.

E Limitations and Future Work

Despite the strong performance of TIGR-Seg, some directions remain for further improvement. First, performance is affected when textual guidance is missing or substantially inconsistent with

Table 9: **Effect of training data size on BKAI and BUSI.** All methods are evaluated on the same test set using Dice and mIoU.

Training Setting	BKAI		BUSI	
	Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
LViT (Li et al., 2023) (100% Training)	85.59	74.81	80.86	67.87
RecLMIS (Huang et al., 2024) (100% Training)	81.44	68.69	82.86	70.74
TIGR-Seg (10% Training)	73.43	59.87	74.51	59.38
TIGR-Seg (25% Training)	82.25	70.61	79.88	66.50
TIGR-Seg (50% Training)	87.25	77.42	83.69	71.95
TIGR-Seg (100% Training)	**92.78**	**86.53**	**86.77**	**76.63**

Table 10: **Effect of image encoders on BKAI and BUSI.** All image encoders are kept frozen.

Encoder	Params (M) ↓	BKAI		BUSI	
		Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
ViT (Dosovitskiy et al., 2021)	3.1	86.24	76.00	79.40	65.84
DeiT (Touvron et al., 2021)	3.1	85.50	74.98	80.55	67.43
SAM2 (Ravi et al., 2025)	4.6	87.03	77.12	82.78	70.62
Swin (Liu et al., 2022)	4.6	**92.78**	**86.53**	**86.77**	**76.63**

the image. Improving robustness to incomplete, noisy, or unreliable textual descriptions therefore remains an important direction. Second, the TIGR Decoder currently uses a single correlated Gaussian as the spatial prior at each decoding stage. While this provides a compact spatial representation and does not constrain the topology of the final segmentation mask, a single Gaussian is less expressive as a spatial prior in extreme multi-focal cases where target regions are widely separated. Adaptive multi-component Gaussian priors could be investigated for such cases. Finally, although TIGR-Seg substantially reduces the number of trainable parameters, the frozen vision and language encoders still contribute to the overall model size. Future work will investigate lighter pretrained encoders and model compression, as well as extension to volumetric medical image segmentation.

Table 11: **Model efficiency comparison.** Trainable parameters, total parameters, FLOPs, model size, and inference latency are reported.

Method	Trainable Params (M) ↓	Total Params (M) ↓	FLOPs (G) ↓	Model Size (MB) ↓	Latency (ms/img) ↓
MADGNet (Nam et al., 2024)	33.4	33.4	8.8	127.2	57.6
MedSA (Wu et al., 2025)	10.6	104.3	55.2	398.0	44.2
EccoMamba (Xu et al., 2026)	38.3	38.3	6.2	242.1	60.1
ReMamber (Yang et al., 2024)	102.5	102.5	7.4	410.4	148.5
ProLearn (Ye et al., 2025)	61.2	257.5	76.3	1028.2	484.62
MedCLIPSeg (Koleilat et al., 2026)	18.7	214.1	73.6	858.3	3503.2
TIGR-Seg (Ours)	**4.6**	**166.9**	**18.5**	**667.9**	**245.4**