DReAM-SAM: Dynamic Region-Adaptive Modulated Segment Anything Model for Fully Automatic Medical Image Segmentation

Kwanseok Oh, *Member, IEEE*, Dahye Lee, and Heung-Il Suk, *Senior Member, IEEE*

***Abstract*— The Segment Anything Model (SAM) has shown strong zero-shot segmentation capabilities across general visual domains. However, its effectiveness in Medical Image Segmentation (MIS) remains constrained by the medical-domain gap, reliance on explicit user prompts, and limited adaptability to small, heterogeneous, or ill-defined targets. In particular, feature adaptation in existing SAM approaches often remains shared across anatomical targets, making it difficult to tailor local representations to target identity, morphology, and surrounding anatomical context. In this paper, we propose Dynamic Region-Adaptive Modulated SAM (DReAM-SAM), a fully automatic framework that introduces target- and region-adaptive representation learning for fine-grained MIS. Our framework introduces three key technical pillars: (1) Target-Conditioned Dynamic Modulation: We propose an agile CNN (ACNN) that complements SAM's encoder by adapting local feature extraction to target identity and morphology while promoting boundary-sensitive details. (2) Query-Adaptive Feature Interaction: We develop a bidirectional neighborhood attention mechanism to integrate complementary ACNN and SAM features over query-adaptive multi-scale contexts. (3) Target-Aware Self-Prompting: We further devise a self-prompting mechanism that exploits target-existence cues and target-aware representations to automatically generate surrogate prompts. Extensive experiments across six MIS datasets and diverse medical imaging modalities reveal that DReAM-SAM outperforms state-of-the-art SAM variants and task-specific MIS models. The code is publicly available at `https://github.com/ku-milab/DReAM-SAM`.**

***Index Terms*— Medical Image Segmentation, Foundation Model Adaptation, Segment Anything Model**

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Department of Artificial Intelligence (Korea University)); the Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents (IITP-2025-RS-2025-02304828) grant funded by the Korea government (MSIT); and the Technology Innovation Program (RS-2025-02221011, Development of Medical-Specialized Multimodal Hyperscale Generative AI Technology for Global Integration) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

Kwanseok Oh is with the Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea and DEEPNOID Inc., Seoul 08377, Republic of Korea (e-mail: ksohh@korea.ac.kr)

Dahye Lee is with DEEPNOID Inc., Seoul 08377, Republic of Korea (e-mail: dhlee@deepnoid.com)

Heung-Il Suk is with the Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea (e-mail: hisuk@korea.ac.kr).

Kwanseok Oh and Dahye Lee contributed equally to this work.

Corresponding author: Heung-Il Suk

I. Introduction

MEDICAL Image Segmentation (MIS) plays a pivotal role in modern healthcare, serving as an initial step in numerous image-driven diagnostic and therapeutic workflows [1]. Recently, deep learning-based methods have become the *de facto* standard for MIS due to their high efficacy and outstanding performance. In this regard, Convolutional Neural Networks (CNNs) [2]–[4], Vision Transformers (ViTs) [5]–[7], and their hybrid architectures [8] have shown remarkable capabilities in learning complex spatial and contextual representations from medical images. Nonetheless, these methods are typically designed and trained for specific MIS tasks, achieving strong performance within their target domains but often requiring extensive training and large-scale annotated datasets to ensure stable convergence.

The emergence of foundation models trained on vast natural image datasets has recently revolutionized the landscape of image segmentation [9]. Among them, the Segment Anything Model (SAM) [10] has garnered significant attention for its general-purpose, prompt-based segmentation capability across diverse visual domains. However, when applied to MIS [11], SAM faces substantial domain gaps inherited from its pretraining on natural scenes and functional limitations arising from its reliance on explicit user prompts. These challenges limit its direct applicability to fully automatic MIS, which requires both medical-domain representations and prompt-free inference.

Our primary goal in this study is to adapt SAM to fully automatic MIS by mitigating the medical-domain gap and eliminating reliance on explicit manual prompts. Existing studies have addressed these challenges from several perspectives. MedSAM [12] and SAM-Med3D [13] adapt SAM to medical domains via large-scale fine-tuning or training on medical images, while parameter-efficient approaches introduce lightweight adapters or low-rank modules [14]–[18] to incorporate domain-specific knowledge. Other studies reduce dependence on manual interaction through automatic or prompt-free strategies [19]–[21], such as cross-prompting, hierarchical decoding, and learned prompt representations. More recently, auxiliary CNN branches have been incorporated alongside SAM [22], [23] to complement its global representations with local image features. While such approaches effectively improve medical-domain adaptation and/or prompt automation, the *feature adaptation operators themselves remain largely*

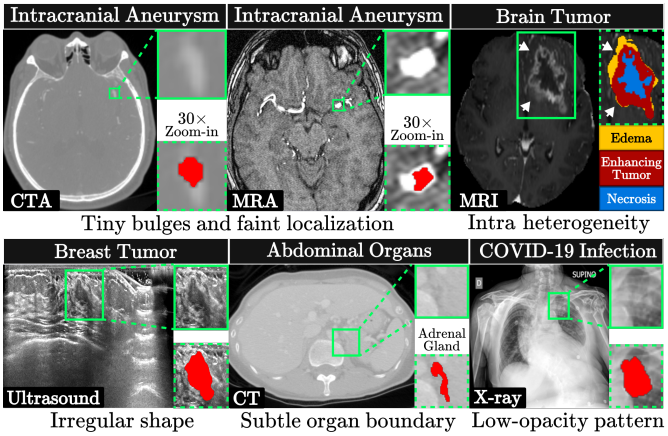

Fig. 1: Various medical imaging modalities present distinct challenges that complicate accurate segmentation.

unconditioned on the semantic identity of the anatomical target being sought. This can be particularly limiting in medical images, where similar local appearances may correspond to different anatomical structures or pathological regions. As a result, target disambiguation is largely deferred to subsequent prediction stages rather than being explicitly incorporated into the feature extraction process. Moreover, CNN-assisted SAM variants [22], [23] typically rely on static convolutional operators and predefined cross-branch interaction patterns, limiting their ability to adapt local representations to target morphology and surrounding anatomical context.

As illustrated in Fig. 1, the need for target-adaptive representation becomes particularly evident in challenging MIS scenarios. Aneurysms in computed tomography angiography (CTA) [24] and magnetic resonance angiography (MRA) [25] often appear as minute vascular bulges with faint localization, whereas brain tumors in magnetic resonance imaging (MRI) [26] exhibit substantial intra-tumoral heterogeneity. Breast lesions in ultrasound [27] frequently present low contrast and ill-defined boundaries due to acoustic shadowing and speckle noise. Similarly, small abdominal organs in computed tomography (CT) [28], such as the adrenal glands or gallbladder, can resemble adjacent tissues, while COVID-19 infections in chest X-ray [29] may manifest as diffuse, low-opacity regions overlapping with surrounding anatomical structures. Importantly, such ambiguities can be difficult to resolve from local appearance alone: visually similar patterns may represent different anatomical structures or heterogeneous pathological sub-regions, depending on the target of interest. This motivates local representations that account not only for *what* is observed at a spatial location, but also for *which target* is being sought and how it relates to the surrounding anatomy. Accordingly, beyond medical-domain adaptation and prompt automation, challenging MIS requires *target-adaptive local representation, boundary-aware learning, and flexible integration of regional anatomical context*.

Motivated by these observations, we propose a **D**ynamic **Re**gion-**A**daptive **M**odulated **SAM** (**DReAM-SAM**), which advances fully automatic SAM adaptation with target- and region-adaptive representation learning. Unlike existing approaches that primarily focus on medical-domain adaptation, prompt automation, or generic local–global feature integration, DReAM-SAM explicitly adapts the feature extraction process according to target-specific cues and spatial context. Specifically, we introduce an agile CNN (ACNN) branch in parallel with SAM's ViT encoder to complement its global representations with target-adaptive local features. Within the ACNN branch, the proposed dynamic region-adaptive block (DyRA-Block) generates target-conditioned convolutional kernels through a shared–specific decomposition, so that anatomical primitives common to all targets are shared while the components distinguishing each target are preserved, together with dedicated regularization for target discriminability and boundary sensitivity. Dual-interactive dilated neighborhood attention (D^2NA) further integrates these fine-grained local features with SAM's global semantics via query-adaptive multi-scale contextual interaction. Finally, the target-existence classifier (TEC) and target-aware prompt generator (TAPG) exploit the resulting target-aware features to generate surrogate prompts without manual intervention. Through this unified design, DReAM-SAM introduces target-conditioned local feature extraction into SAM adaptation while preserving SAM's global semantic capability and enabling fully automatic MIS under full supervision with dataset-specific target definitions.

The main contributions are summarized as follows:

- We propose DReAM-SAM, a fully automatic SAM adaptation framework that explicitly introduces target-/region-adaptive representation learning to address the lack of target-adaptive local representation in challenging MIS.
- We introduce an ACNN branch equipped with the DyRA-Block, which performs target-conditioned dynamic feature extraction via a shared–specific kernel decomposition and spatially adaptive sampling, complemented by kernel decorrelation and boundary-aware region decoupling.
- We further develop D^2NA for query-adaptive multi-scale interaction between the complementary ACNN and SAM representations; and a self-prompting module (TEC + TAPG) that derives surrogate prompts from inferred target cues without manual intervention.
- Extensive experiments across six MIS datasets and diverse imaging modalities demonstrate that DReAM-SAM outperforms state-of-the-art task-specific and SAM-based segmentation approaches, with particularly pronounced gains on small, heterogeneous, and ambiguous targets.

II. Related Works

A. CNN-/ViT-Based MIS Methods

Over the past decade, CNNs have remained the cornerstone of MIS. In particular, U-Net [2] and its variants popularized the encoder–decoder structure with skip connections to preserve spatial detail. Among them, nnUNet [3] extended this idea into a self-configuring framework that automatically adapts to different datasets, while MedNeXt [4] modernized CNN-based segmentation using ConvNeXt-style [30], [31] convolutional blocks for improved efficiency and scalability. Another research stream has refined or enforced CNN blocks by exploiting deformable convolution [32]–[36], neighborhood attention

TABLE I: Mechanism-level comparison of recent SAM-based methods for MIS. "Target-identity conditioning" indicates whether target/task identity is introduced at the prompt level or directly into convolutional kernels. "Conditioning regularization" indicates whether an objective keeps such conditioned components mutually distinguishable. "Cross-branch attention scope" denotes the region each query attends to in the other branch: the entire feature map (*global*), the footprint of the query position only (*co-located*), or dilated neighborhoods weighted per query (*query-adaptive multi-scale*).

Methods	Adaptation	Convolutional kernels	Target-identity conditioning	Conditioning regularization	Boundary supervision	Cross-branch attention scope	Prompt source
MedSAM	Full fine-tuning	–	✗	–	✗	–	Manual (box)
3DSAM-adt	Adapter	–	✗	–	✗	–	Manual (point)
CPC-SAM	Dual decoder	–	✗	–	✗	–	Dual-decoder prompting
H-SAM	LoRA + decoder	–	✗	–	✗	–	Prior mask (prompt-free)
DAPSAM	Adapter	–	✗	–	✗	–	Prototype memory bank
Med-SA	Adapter	–	✗	–	✗	–	Manual (point)
AutoSAMUS	Adapter + CNN	Static	✗	–	✗	Global, single-scale	Image embedding
SAMCT	Adapter + CNN	Static	Prompt-level	✗	✗	Co-located window, single-scale	Task indicator
DReAM-SAM	**LoRA + CNN**	**Dynamic (shared-specific)**	**Kernel-level**	**Decorrelation**	**Region decoupling**	**Neighborhood, query-adaptive multi-scale**	**Self-prompting from inferred target cues**

(**Note**) "✗" marks an absent mechanism; "–" marks an inapplicable axis (no auxiliary branch or no target conditioning)

(NA) [37] variants [38], [39], and large-kernel designs [40]–[42]. Meanwhile, ViT-based methods such as SwinUNETR [6] and nnFormer [7] have been proposed to capture long-range dependencies via volumetric self-attention mechanisms. As a hybrid design, TransUNet [8] integrates ViT into a U-Net backbone to jointly leverage both local and global features.

B. Adapting SAM for Medical Image Segmentation

The remarkable versatility and generalization capability of SAM [10] have led to revolutionary changes in the field of MIS. A straightforward strategy for adapting SAM to medical domains involves directly updating on large-scale medical datasets: MedSAM [12] and SAM-Med3D [13]. Beyond such data-driven adaptations, several studies have explored parameter-efficient or architecture-level modulations. SAM-Med2D [15] and SAMed [17] employ low-rank decomposition (*e.g.*, LoRA [43] or FacT [44]) for efficient fine-tuning, while 3DSAM-adt [14], MA-SAM [16], and Med-SA [18] integrate lightweight adapters to infuse domain-specific knowledge without full parameter updates. In parallel, several works have focused on automatic prompt generation to mitigate reliance on manual inputs. These include prototype-based prompt construction using a learnable memory bank (DAPSAM [45]), cross-prompting via dual-decoder architectures (CPC-SAM [19]) and its synergistic training variant (SynFoC [20]), as well as hierarchical decoding mechanisms that progressively refine segmentation results (H-SAM [21]). Other research directions further seek to enhance SAM's capacity to capture local features by introducing auxiliary networks. AutoSAMUS [22] and SAMCT [23] utilize an additional CNN branch parallel to the SAM encoder, enabling the integration of local and global contextual features through cross-branch interactions.

Tab. I presents a mechanism-level comparison of representative SAM-based approaches for MIS. Existing methods primarily focus on medical-domain adaptation, prompt automation, or auxiliary local–global feature integration, while target or task cues are generally absent from feature extraction or introduced only at the prompt level. Moreover, CNN-assisted variants [22], [23] typically rely on static convolutional operators shared across targets and fixed cross-branch interaction scopes. In contrast, the proposed DReAM-SAM conditions local feature extraction on target identity, keeps the resulting target-specific components distinguishable through explicit decorrelation, and adapts contextual interaction to each spatial query, thereby yielding target-aware representations that are further leveraged for automatic prompt generation.

III. Method

As illustrated in Fig. 2, the proposed DReAM-SAM is composed of the original SAM augmented with an ACNN branch, D^{2}NA blocks, and a self-prompting module consisting of a target-existence classifier (TEC) and TAPG.

The SAM branch follows the standard SAM image encoder $\mathcal{E}_{\text{s}}$, where an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ is first tokenized via patch embedding and then processed by ViT encoder blocks to extract rich global semantics, with H and W denoting the spatial dimensions. In parallel, the ACNN branch is designed to capture fine-scale anatomical details that complement the SAM features. It begins with a stem block followed by a series of vanilla CNN blocks and culminates in the proposed DyRA-Block (see Fig. 3), which partitions input features into target-decoupled attention maps. For each target, the DyRA-Block dynamically generates spatial offsets and target-conditioned kernels via a shared-specific decomposition, enabling morphology- and scale-aware feature modulation. To effectively integrate heterogeneous representations from both branches, we introduce the D^{2}NA block (see Fig. 4), which comprises two reconciliation streams: self-DNA and cross-DNA. The self-DNA stream refines each branch's features through locally adjacent aggregation, while the cross-DNA stream promotes bidirectional interaction between the SAM and ACNN branches, harmonizing global semantics with localized anatomical cues. Finally, DReAM-SAM incorporates a self-prompting mechanism to replace manual prompt inputs. The TEC predicts the presence of target classes and provides corresponding semantic cues to the TAPG, which, conditioned on these cues and decoupled target attention-based guidance, yields target-specific surrogate prompts. These prompts, together with the refined SAM branch's visual representations,

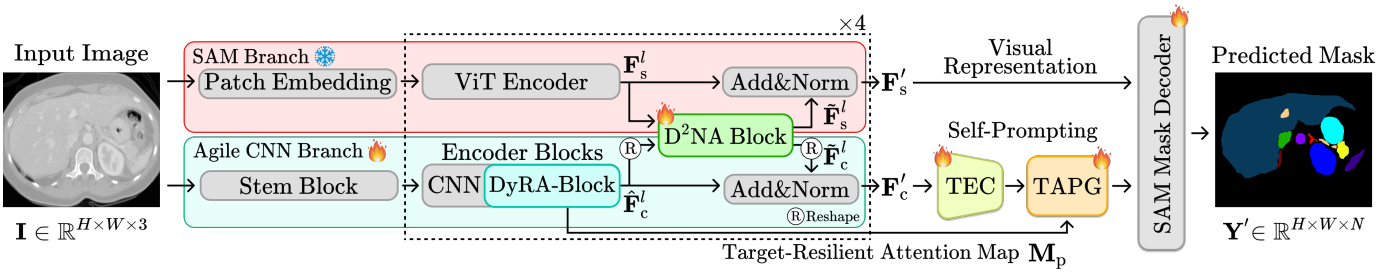

Fig. 2: Schematic overview of the proposed DReAM-SAM. It comprises three key strategies: (i) DyRA-Block; (ii) D^{2}NA block; and (iii) Self-prompting module (TEC + TAPG).

are fed into the SAM mask decoder $\mathcal{D}$ to produce accurate segmentation masks $\mathbf{Y}' \in \mathbb{R}^{H\times W\times N}$.

A. Agile CNN Encoder Branch

The ACNN branch is designed to complement the global semantics with fine-grained anatomical details via dynamically target-adaptive convolutional modeling. It comprises three major components: (i) a target region decoupler that isolates target-specific attention maps, (ii) a target-conditioned kernel generator that produces adaptive convolutional kernels via a shared-specific decomposition, and (iii) an offset generator that enables spatially flexible sampling. Together, these components constitute the proposed DyRA-Block, which enhances the spatial adaptability and target sensitivity of CNN features. Given an input image $\mathbf{I}$, the vanilla CNN encoder block $\mathcal{E}_\text{c}$ extracts multi-level features $\{\mathbf{F}_\text{c}^l\}_{l=1}^L = \mathcal{E}_\text{c}(\mathbf{I})$, where L is the total number of convolutional stages. Without loss of generality, the following formulations describe the operations at the l-th feature level $\mathbf{F}_\text{c}^l$.

1) Target-region decoupler: To effectively capture discriminative and region-aware representations, the DyRA-Block (see Fig. 3) first provides the target-decoupled attention maps $\mathbf{M}^l \in \mathbb{R}^{H'\times W'\times N}$ from the input features $\mathbf{F}_\text{c}^l$ via the 1×1 convolution ($\text{Conv}_{1\times 1}$) followed by a Softmax activation as:

$$\mathbf{M}^l = \text{Softmax}\left(\text{Conv}_{1\times 1}\left(\mathbf{F}_\text{c}^l\right)\right), \tag{1}$$

where N is the total number of targets. Each target-specific attention map $\mathbf{M}_n^l$ highlights salient regions and is applied to $\mathbf{F}_\text{c}^l$ to yield target-attended features:

$$\bar{\mathbf{F}}_n^l = (\mathbf{M}_n^l \otimes \mathbf{F}_\text{c}^l) \oplus \mathbf{F}_\text{c}^l, \tag{2}$$

where $\otimes$ and $\oplus$ are element-wise multiplication and addition. To align such target-wise assignment with the actual anatomy, $\mathbf{M}^l$ is explicitly supervised at every stage so that each target region is decoupled from its surroundings, with emphasis near ambiguous boundaries (refer to Sec. III-D.4).

2) Target-conditioned kernel generator: Unlike prior dynamic convolutions [32], [46] that derive kernels solely from input features, we additionally condition the kernel on the *identity* of the target being processed. This is motivated by the observation that visually similar regions in medical images often belong to different anatomical structures (*e.g.*, adjacent abdominal organs or co-existing tumor sub-regions), so identical local appearance demands different filtering depending on which target is being sought. Yet, anatomical structures

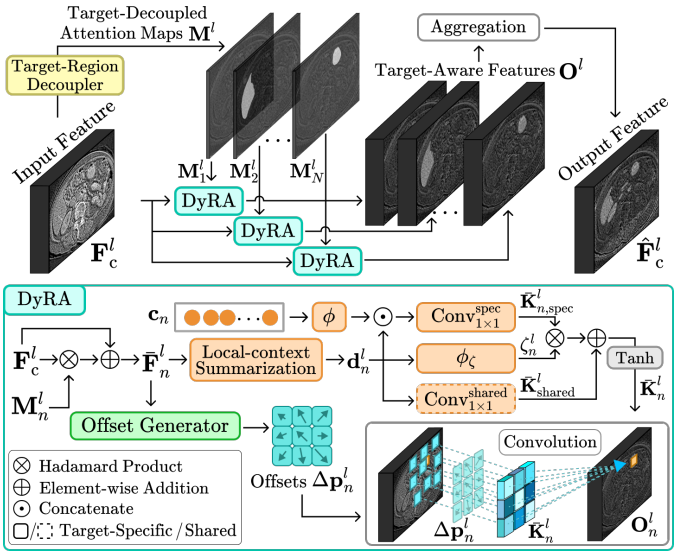

Fig. 3: Illustration of the DyRA-Block, which generates target-conditioned kernels via a shared-specific decomposition and applies them at offset-shifted positions for spatially adaptive, target-sensitive filtering.

share many low-level visual primitives, such as locally homogeneous textures, and forcing every target to learn these patterns independently would be both redundant and prone to overfitting under limited supervision. To reconcile these competing demands, we decompose each target-conditioned kernel into a *shared feature basis* and a *target-specific residual*, capturing what is common across targets while preserving target-dependent characteristics.

Spatial responses of $\bar{\mathbf{F}}_n^l$ from Eq. (2) are first condensed via average pooling (AvgPool) into a kernel descriptor: $\mathbf{K}_n^l = \text{AvgPool}(\bar{\mathbf{F}}_n^l) \in \mathbb{R}^{K\times K\times C}$, where $K\times K$ represents the reduced spatial resolution and C is the channel dimension. $\mathbf{K}_n^l$ is then encoded by a $\text{Conv}_{3\times 3}$–ReLU block followed by global average pooling (GAP), yielding a channel-wise descriptor $\mathbf{d}_n^l \in \mathbb{R}^C$ that summarizes the local spatial context. To inject target-specific identity, we introduce a set of learnable target prototypes $\{\mathbf{c}_n\}_{n=1}^N \in \mathbb{R}^C$, with each $\mathbf{c}_n$ serving as a location-agnostic embedding that captures target-discriminative characteristics learned during training. A target-specific kernel $\bar{\mathbf{K}}_{n,\text{spec}}^l$ is produced by channel-wise concatenating $\mathbf{d}_n^l$ with the projected target prototype $\phi(\mathbf{c}_n) \in \mathbb{R}^C$ and expanding the

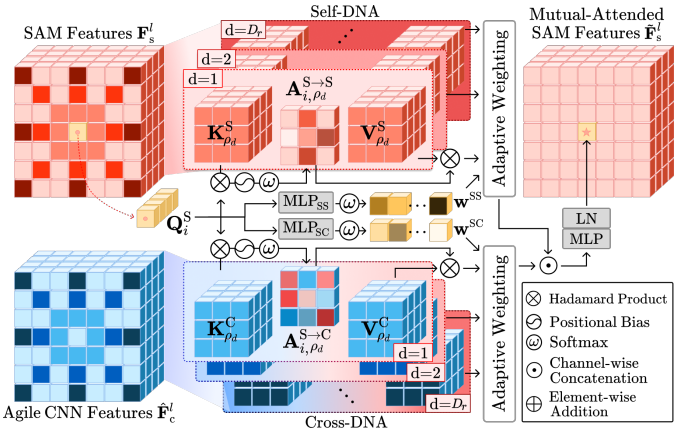

Fig. 4: Illustration of D^{2}NA, consisting of self-DNA for intra-branch contextual aggregation and cross-DNA for inter-branch feature interaction with adaptive multi-dilation weighting.

result via $\mathrm{Conv}_{1\times1}^{\mathrm{spec}}$ to $\mathbb{R}^{(C\cdot\bar{K}\cdot\bar{K})\times1\times1}$:

$$\bar{\mathbf{K}}_{n,\mathrm{spec}}^l = \mathrm{Reshape}\left(\mathrm{Conv}_{1\times1}^{\mathrm{spec}}(\mathbf{d}_n^l \odot \phi(\mathbf{c}_n))\right) \in \mathbb{R}^{\bar{K}\times\bar{K}\times C}, \quad (3)$$

where $\phi(\cdot)$ denotes a lightweight projection layer. In parallel, a target-agnostic shared kernel $\bar{\mathbf{K}}_{\mathrm{shared}}^l \in \mathbb{R}^{\bar{K}\times\bar{K}\times C}$ is generated from $\mathbf{d}_n^l$ alone via a separate weight-shared $\mathrm{Conv}_{1\times1}^{\mathrm{shared}}$, capturing common feature patterns shared across all N targets at stage l. The two kernels are then composed into the final target-conditioned kernel $\bar{\mathbf{K}}_n^l$:

$$\bar{\mathbf{K}}_n^l = \mathrm{Tanh}\left(\bar{\mathbf{K}}_{\mathrm{shared}}^l + \zeta_n^l \cdot \bar{\mathbf{K}}_{n,\mathrm{spec}}^l\right), \quad (4)$$

where $\zeta_n^l = \mathrm{Sigmoid}(\phi_\zeta(\mathbf{d}_n^l))$ denotes a learned scalar gate that modulates how strongly the target-specific component contributes for each target and stage. Consequently, $\bar{\mathbf{K}}_n^l$ is designed to encode both *what is shared across targets* and *what distinguishes* each of them, with the gate ζ_n^l adapting the balance per target and per scale.

3) Offset-guided adaptive convolution: For spatially adaptive sampling, we introduce an offset generator that predicts per-location displacements conditioned on the target-attended features. Given $\bar{\mathbf{F}}_n^l$, a lightweight stack of two $\mathrm{Conv}_{3\times3}$ layers with intermediate ReLU activation produces the offset field $\Delta\mathbf{p}_n^l = \mathrm{Conv}_{\mathrm{offset}}(\bar{\mathbf{F}}_n^l) \in \mathbb{R}^{\bar{K}\times\bar{K}\times2}$, where each entry encodes a 2D displacement $(\delta x, \delta y)$ for the corresponding kernel position. The target-conditioned kernels $\bar{\mathbf{K}}_n^l \in \mathbb{R}^{\bar{K}\times\bar{K}\times C}$ derived from Eq. (4) are subsequently applied at offset-shifted positions via deformable convolution $(*)$ [32]:

$$\mathbf{O}_n^l(\mathbf{p}_0) = \sum_{\mathbf{k}\in\mathcal{R}_k} \bar{\mathbf{K}}_n^l(\mathbf{k}) * \bar{\mathbf{F}}_n^l\left(\mathbf{p}_0 + \mathbf{k} + \Delta\mathbf{p}_n^l(\mathbf{k})\right), \quad (5)$$

where $\mathbf{p}_0$ denotes a reference spatial location, $\mathcal{R}_k$ represents the regular $\bar{K}\times\bar{K}$ kernel grid, and $\mathbf{k}\in\mathcal{R}_k$ indexes each kernel position with its predicted offset $\Delta\mathbf{p}_n^l(\mathbf{k})$.

Finally, target-aware features $\mathbf{O}^l$ from all N decoupled regions are incorporated via channel-wise stacking $(\|)$ and fused through a $\mathrm{Conv}_{1\times1}$ to yield the refined output as $\hat{\mathbf{F}}_c^l := \mathrm{Conv}_{1\times1}\left(\|_{n=1}^N \mathbf{O}_n^l\right) \in \mathbb{R}^{H'\times W'\times C}$.

B. Dual-Interactive Dilated Neighborhood Attention

In Fig. 4, D^{2}NA is designed to facilitate bidirectional feature interaction between the SAM and ACNN branches while efficiently capturing multi-scale contextual dependencies. Building on NA [37], which captures spatial locality via restricted attention windows, we extend this concept with (i) a self-DNA that aggregates contextual cues from spatially sparse yet semantically relevant dilated neighbors, and (ii) a cross-DNA that bidirectionally exchanges such cues with the other branch. This enables the SAM branch to benefit from fine-grained local details, while the ACNN branch gains global semantic awareness, allowing mutually reinforced feature learning.

1) Dilated neighborhood definition: For a reference token index i corresponding to the 2D spatial position, the dilated neighborhood formula is defined as $\rho_d(i) = \{i + d\cdot\delta \mid \delta \in \mathcal{R}_{k_d\times k_d}\}$, where d denotes the dilation factor, and $\mathcal{R}_{k_d\times k_d}$ is the kernel grid for dilation. Increasing d enlarges the spatial sampling interval among neighboring tokens while preserving the kernel size k_d, allowing the model to capture broader contextual information with minimal computation.

2) Self-DNA: Taking the SAM branch as an example (see Fig. 4), feature maps $\{\mathbf{F}_s^l\}_{l=1}^L = \mathcal{E}_s(\mathbf{I})$ are first reshaped into the sequence tokens N_T as $\mathbf{F}_s^l \in \mathbb{R}^{N_T\times C'}$, and then query, key, and value embeddings are obtained as $\mathbf{Q}^{\mathrm{S}} = \mathbf{F}_s^l\mathbf{W}_{Q_s}, \mathbf{K}^{\mathrm{S}} = \mathbf{F}_s^l\mathbf{W}_{K_s}, \mathbf{V}^{\mathrm{S}} = \mathbf{F}_s^l\mathbf{W}_{V_s}$, where $\mathbf{W}_{Q_s,K_s,V_s} \in \mathbb{R}^{C'\times D}$ are learnable projection matrices. For each query token $\mathbf{Q}_i^{\mathrm{S}}$, the self-attention weights $\mathbf{A}_{i,\rho_d}^{\mathrm{S}\to\mathrm{S}}$ to its dilated neighbors $\rho_d(i)$ are computed as:

$$\mathbf{A}_{i,\rho_d}^{\mathrm{S}\to\mathrm{S}} = \left[\frac{\exp((\mathbf{Q}_i^{\mathrm{S}}\mathbf{K}_{j_d}^{\mathrm{S}\top})/\eta + B_{(i,j_d)})}{\sum_{t_d\in\rho_d(i)}\exp((\mathbf{Q}_i^{\mathrm{S}}\mathbf{K}_{t_d}^{\mathrm{S}\top})/\eta + B_{(i,t_d)})}\right]_{j_d\in\rho_d(i)}, \quad (6)$$

where η and $B_{(i,\cdot)}$ denote the temperature scaling and learnable relative positional bias, respectively. To effectively consider multi-scale contextual knowledge, responses from multiple dilation factors $\{1,\ldots,D_r\}$ are fused using query-dependent weights:

$$\mathbf{a}_i^{\mathrm{SS},l} := \sum_{d=1}^{D_r} \mathbf{w}^{\mathrm{SS}}[d]\cdot\left(\mathbf{A}_{i,\rho_d}^{\mathrm{S}\to\mathrm{S}} \otimes \mathbf{V}_{\rho_d}^{\mathrm{S}}\right), \quad (7)$$

where $\mathbf{w}^{\mathrm{SS}} = \mathrm{Softmax}(\mathrm{MLP}_{\mathrm{SS}}(\mathbf{Q}_i^{\mathrm{S}})) \in \mathbb{R}^{1\times D_r}$ emphasizes the most informative receptive field for each query position i. A symmetric process in Eqs. (6), (7) is applied to the ACNN branch features $\hat{\mathbf{F}}_c^{l*}$ to yield the self-attention output $\mathbf{a}_i^{\mathrm{CC},l}$.

3) Cross-DNA: In this stream, the query originates from one branch while the key-value pairs are derived from the other, enabling the transfer of complementary cues. For instance, SAM-to-ACNN cross-attention is formulated as:

$$\mathbf{A}_{i,\rho_d}^{\mathrm{S}\to\mathrm{C}} = \left[\frac{\exp((\mathbf{Q}_i^{\mathrm{S}}\mathbf{K}_{j_d}^{\mathrm{C}\top})/\eta + B_{(i,j_d)})}{\sum_{t_d\in\rho_d(i)}\exp((\mathbf{Q}_i^{\mathrm{S}}\mathbf{K}_{t_d}^{\mathrm{C}\top})/\eta + B_{(i,t_d)})}\right]_{j_d\in\rho_d(i)}, \quad (8)$$

where $\mathbf{K}^{\mathrm{C}} = \hat{\mathbf{F}}_c^l\mathbf{W}_{K_c}$ and $\mathbf{V}^{\mathrm{C}} = \hat{\mathbf{F}}_c^l\mathbf{W}_{V_c}$. The aggregated attention response is subsequently computed as:

$$\mathbf{a}_i^{\mathrm{SC},l} := \sum_{d=1}^{D_r} \mathbf{w}^{\mathrm{SC}}[d]\cdot\left(\mathbf{A}_{i,\rho_d}^{\mathrm{S}\to\mathrm{C}} \otimes \mathbf{V}_{\rho_d}^{\mathrm{C}}\right), \quad (9)$$

*For the ACNN branch, a reshape is applied before and after the D^{2}NA block to ensure spatial alignment with the corresponding $\mathbf{F}_s^l$ (see Fig. 2).

where $\mathbf{w}^{\mathrm{SC}} = \mathrm{Softmax}(\mathrm{MLP}_{\mathrm{SC}}(\mathbf{Q}_i^{\mathrm{S}}))$. For the opposite scenario (*i.e.*, ACNN-to-SAM), the identical operation is symmetrically applied to yield $\mathbf{a}_i^{\mathrm{CS},l}$.

4) Feature fusion: The outputs from self-DNA and cross-DNA streams are concatenated channel-wise ($\odot$), followed by the MLP layer with layer normalization (LN). The resulting token-wise feature vectors are subsequently reassembled into their original arrangement to reconstruct full 2D feature maps:

$$\begin{aligned}\tilde{\mathbf{F}}_{\mathrm{s}}^{l} &= \Big\|_{i=1}^{N_T} \mathrm{LN}(\mathrm{MLP}(\mathbf{a}_i^{\mathrm{SS},l} \odot \mathbf{a}_i^{\mathrm{SC},l})),\\ \tilde{\mathbf{F}}_{\mathrm{c}}^{l} &= \Big\|_{i=1}^{N_T} \mathrm{LN}(\mathrm{MLP}(\mathbf{a}_i^{\mathrm{CC},l} \odot \mathbf{a}_i^{\mathrm{CS},l})).\end{aligned} \tag{10}$$

Here, $\|$ stacks the processed feature vectors of all N_T tokens according to their spatial positions. The outputs are finally combined with their corresponding branch features via residual connections to preserve branch-specific information: $\mathbf{F}'_{\mathrm{s}} = \mathrm{LN}(\tilde{\mathbf{F}}_{\mathrm{s}}^{l} + \mathbf{F}_{\mathrm{s}}^{l})$, $\mathbf{F}'_{\mathrm{c}} = \mathrm{LN}(\tilde{\mathbf{F}}_{\mathrm{c}}^{l} + \hat{\mathbf{F}}_{\mathrm{c}}^{l})$.

C. Target-Aware Prompt Generation

This module enables DReAM-SAM to autonomously produce semantic prompts without manual intervention, which is composed of (i) a TEC that estimates target-specific priors from the ACNN branch, and (ii) a TAPG that generates surrogate prompts conditioned on these priors.

1) TEC: Given accentuated ACNN features $\mathbf{F}'_{\mathrm{c}}$, TEC predicts class-specific existence probabilities of the current input slice, which serve as semantic priors for prompt modulation. Specifically, $\mathbf{F}'_{\mathrm{c}}$ is globally averaged and passed via an MLP classifier f followed by a Sigmoid activation: $\mathbf{y}' = \mathrm{Sigmoid}(f(\mathrm{GAP}(\mathbf{F}'_{\mathrm{c}}))) \in \mathbb{R}^{1\times N}$.

2) Dense prompt generation: To generate pixel-level surrogate prompts, multiscale attention maps $\{\mathbf{M}^l\}_{l=1}^{L}$ derived from all stages L (refer to Eq. (1)) are resized via linear interpolation $g(\cdot)$ and concatenated along the channel dimension. The target-resilient attention map $\mathbf{M}_{\mathrm{p}}$ is obtained after $\mathrm{Conv}_{1\times1}$ as $\mathbf{M}_{\mathrm{p}} = \mathrm{Conv}_{1\times1}\left(\|_{l=1}^{L} g\left(\mathbf{M}^l\right)\right)$. The resulting $\mathbf{M}_{\mathrm{p}}$ is modulated by target-specific priors $\mathbf{y}'$ and passed through a refinement block ($\mathrm{Conv}_{\mathrm{ref}}$: $\mathrm{Conv}_{3\times3}$–LN–GELU–$\mathrm{Conv}_{1\times1}$) to produce dense prompts: $\mathbf{P}_{\mathrm{d}} = \mathrm{Conv}_{\mathrm{ref}}(\mathbf{M}_{\mathrm{p}} \otimes \mathbf{y}')$.

3) Sparse prompt generation: In parallel, TAPG produces sparse prompts that encode compact semantic cues. The $\mathbf{M}_{\mathrm{p}}$ is flattened and projected via an MLP with Tanh activation as $\mathbf{m}_{\mathrm{p}} = \mathrm{Tanh}(\mathrm{MLP}(\mathrm{Flatten}(\mathbf{M}_{\mathrm{p}}))) \in \mathbb{R}^{1\times D}$. To further incorporate class-guided information, a static descriptor $\mathbf{C}_{\mathrm{des}} \in \mathbb{R}^{2\times D}$ is used as a routing signal for target existence: $\mathbf{P}_{\mathrm{s}} = \mathbf{m}_{\mathrm{p}} + \|_{n=1}^{N} \mathbf{C}_{\mathrm{des}}[\mathbb{1}\{\mathbf{y}'[n] \geq \tau\}, :]$ where $\mathbb{1}\{\cdot\}$ is an indicator function and τ is a confidence threshold for selecting target class-relevant descriptors. Here, $\mathbf{C}_{\mathrm{des}}$ is randomly initialized and fixed during training to provide a stable reference.

Accordingly, the SAM decoder $\mathcal{D}$ takes the accentuated SAM features $\mathbf{F}'_{\mathrm{s}}$ with surrogate prompts to produce the segmentation output $\mathbf{Y}' := \mathcal{D}(\mathbf{F}'_{\mathrm{s}}, \mathbf{P}_{\mathrm{d}}, \mathbf{P}_{\mathrm{s}})$.

D. Objective Functions

1) Segmentation loss: We combine Dice and cross-entropy (CE) losses between the predicted mask $\mathbf{Y}'$ and the ground-truth (GT) mask $\mathbf{Y}$ as $\mathcal{L}_{\mathrm{seg}} = Dice(\mathbf{Y}', \mathbf{Y}) + CE(\mathbf{Y}', \mathbf{Y})$.

2) Classification loss: To endow reliable class-level priors for TAPG, the TEC is supervised by binary CE (BCE) on $\mathbf{y}'$ using the slice-level target class label $\mathbf{y}$, derived from the corresponding GT mask $\mathbf{Y}$, as $\mathcal{L}_{\mathrm{cls}} = BCE(\mathbf{y}', \mathbf{y})$.

3) Kernel decorrelation loss: To prevent the target-specific residuals $\bar{\mathbf{K}}_{n,\mathrm{spec}}^{l}$ from collapsing into similar patterns across targets, we introduce a kernel decorrelation (KD) loss that encourages them to be mutually distinguishable:

$$\mathcal{L}_{\mathrm{kd}} = \tfrac{1}{L}\textstyle\sum_{l=1}^{L} \frac{1}{N(N-1)} \sum_{n\neq m} \omega_{nm} \cdot \left|\langle \tilde{\mathbf{K}}_{n,\mathrm{spec}}^{l}, \tilde{\mathbf{K}}_{m,\mathrm{spec}}^{l} \rangle\right|, \tag{11}$$

where $\tilde{\mathbf{K}}_{n,\mathrm{spec}}^{l}$ is the ℓ_2-normalized $\bar{\mathbf{K}}_{n,\mathrm{spec}}^{l}$, and $\omega_{nm} = 1 - \mathrm{Sigmoid}(\mathbf{c}_n^{\top}\mathbf{c}_m/\tau_c)$ softens the penalty for semantically related pairs while enforcing orthogonality for unrelated ones.

4) Region decoupling loss: To explicitly supervise the generation of decoupled attention maps $\mathbf{M}^l$ (Eq. (1)), we introduce a region decoupling (RD) loss that enforces clear separation of target regions. The mask GT $\mathbf{Y}$ is first dilated with radius r_d and blurred by a Gaussian kernel to form a soft target $\widehat{\mathbf{Y}} = \mathrm{Blur}(\mathrm{Dilate}(\mathbf{Y}, r_d))$. To align with the spatial scale of each convolutional stage l, a resizing operator $g(\cdot, l)$ interpolates $\widehat{\mathbf{Y}}$ to each resolution of $\mathbf{M}^l$. The RD loss is subsequently formulated with a weighted pixel-wise discrepancy:

$$\mathcal{L}_{\mathrm{rd}} = \frac{1}{L}\sum_{l=1}^{L} \frac{1}{|\Omega|} \sum_{z\in\Omega} \beta^l(z) \left|\mathbf{M}^l(z) - g(\widehat{\mathbf{Y}}(z), l)\right|, \tag{12}$$

where Ω is the spatial lattice of the image, and $z \in \Omega$ indexes each pixel. The boundary-aware weighting $\beta^l(z)$ is derived from a signed distance transform $dist(g(\mathbf{Y}, l))$: $\beta^l(z) = \exp\left(-\frac{dist(g(\mathbf{Y},l))(z)^2}{2\gamma^2}\right) + \epsilon$, where γ controls the effective boundary band, and ϵ prevents vanishing weights.

5) Overall objective: The overall training objective of DReAM-SAM is defined as: $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{seg}} + \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{kd}} + \lambda \cdot \mathcal{L}_{\mathrm{rd}}$, where λ denotes the weighting coefficient.

IV. Experiments

A. Datasets

1) LIASD [24]: comprises 1,338 CTA images with a total of 1,489 aneurysms (both unruptured and ruptured) collected from six institutions. Following the original protocol in [24], the dataset is split into 1,186 training images and 152 test images, with the test set including 50 normal cases.

2) ADAM [25]: provides MRA scans from 113 subjects (93 patients with 129 unruptured aneurysms, 20 normal). Following the ADAM challenge protocol [25], only untreated cases are considered. In this experiment, we randomly split the data 7:3, yielding 79 training and 34 test samples, of which 6 test samples contain no aneurysm.

3) BTCV [28]: is a benchmark abdominal CT dataset widely used for multi-organ segmentation. It contains 30 volumetric CT scans annotated for 13 abdominal organs. The dataset is split into training and test sets at a 7:3 ratio, with 21 samples in the training set and 9 samples in the test set.

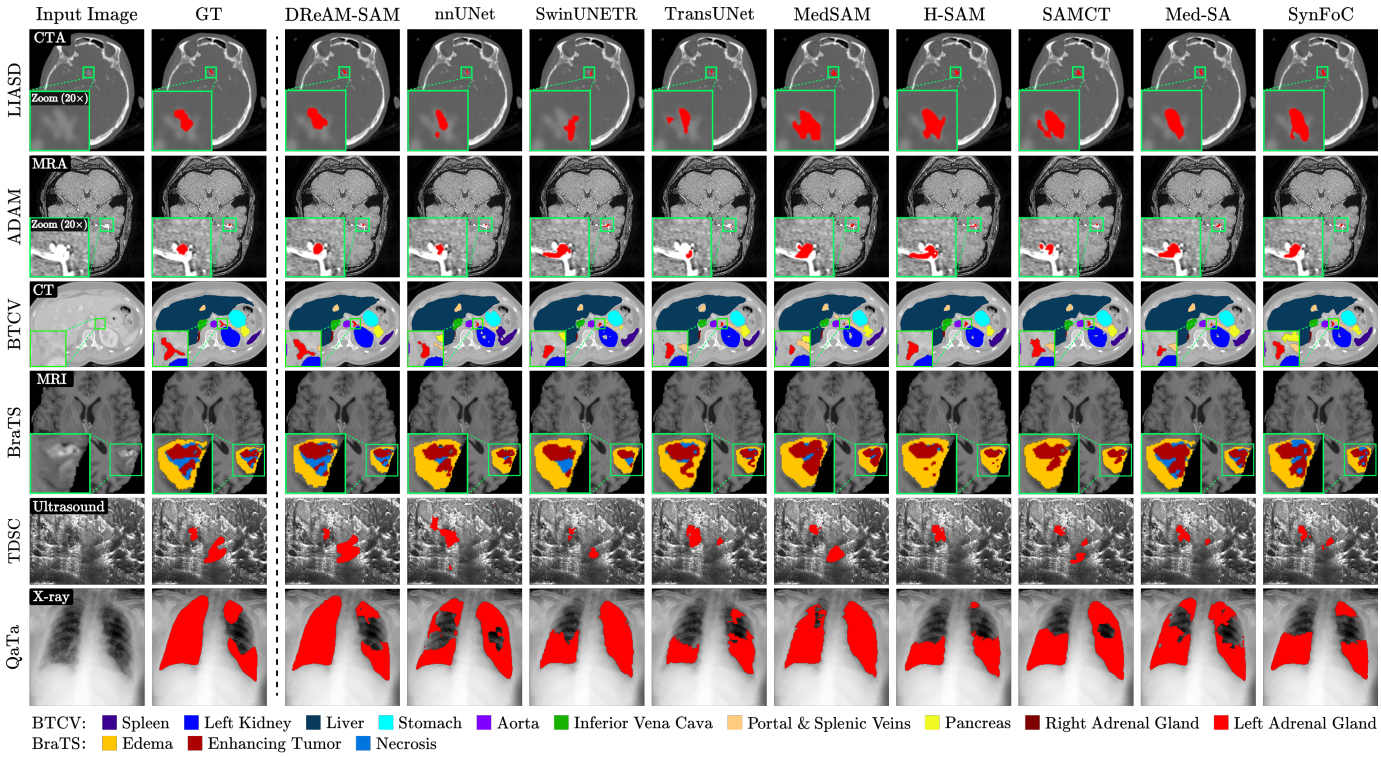

Fig. 5: Qualitative comparison of segmentation results across six challenging datasets.

4) BraTS2021 [26]: consists of 1,251 subjects diagnosed with brain tumors, each with four 3D MRI modalities: native T1, post-contrast T1-weighted (T1CE), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). Following [26], the annotated tumor subregions are grouped into three classes: enhancing tumor (ET), tumor core (TC), and whole tumor (WT). In our experiments, the original training dataset is randomly divided into a train and test set in a 7:3 ratio, yielding 875 and 376 subjects, respectively.

5) TDSC-ABUS [27]: is a breast cancer screening dataset collected from the Harbin Medical University Cancer Hospital using the Invenia ABUS system. It contains 100 3D ultrasound volumes (58 malignant and 42 benign) with expert-annotated tumor boundaries. Among them, 70 volumes are used for training and 30 for testing.

6) QaTa-COV19 [29]: is a large-scale chest X-ray dataset for COVID-19 segmentation. In this study, we utilize v2 subset (*i.e.*, QaTa-COV19-v2), which contains 9,258 pixel-annotated images for pneumonia regions and includes an official split of 7,145 training and 2,113 test images.

B. Implementation Details

Although DReAM-SAM is primarily described based on the 2D SAM backbone in Sec. III, we further demonstrate its flexibility by replacing the backbone with MedSAM [12] and SAM-Med3D [13], enabling evaluation across both 2D and 3D segmentation paradigms within a unified framework. All datasets are preprocessed according to the input specifications of each backbone to ensure consistency across configurations. Unless otherwise specified, DReAM-SAM$^{\text{SAM (2D)}}$ (*i.e.*, **Ours**$^{\text{SAM (2D)}}$) is used as **default** configuration.

- **2D Backbones (SAM / MedSAM):** For experiments using SAM [10] or MedSAM [12] as the backbone, volumetric scans are processed slice-by-slice, where all constituent slices are used regardless of target presence. Each slice is resampled in-plane to $1024 \times 1024 \times 3$ to match the native input resolution of SAM-based encoders, while the original inter-slice spacing is preserved along the depth axis. Intensity values are then standardized via per-volume z-score normalization.
- **3D Backbone (SAM-Med3D):** For experiments with SAM-Med3D [13], we adopt a patch-based volumetric pipeline. All volumes are first resampled to an isotropic spacing of $1.5 \times 1.5 \times 1.5$ mm to ensure spatial consistency across datasets, followed by per-volume z-score normalization. During training, a single $128 \times 128 \times 128$ patch is randomly cropped from each volume. In inference, full-volume predictions are obtained via a sliding-window strategy with overlapping patches.

C. Training Configurations and Evaluation Protocols

All experiments, including those for the competing methods and DReAM-SAM, are conducted in the same PyTorch environment on a single NVIDIA Quadro RTX 8000 (48 GB). For DReAM-SAM, optimization is performed with AdamW ($\beta = (0.9, 0.999)$), an initial learning rate of 1×10^{-5}, and a weight decay of 1×10^{-3}. The model is trained for 50 epochs with a batch size of 4. For the SAM branch, we adopt ViT-B initialized from the released SAM weights and employ LoRA [43] ($r = 4$) for parameter-efficient fine-tuning. The complementary ACNN branch (Sec. III-A) is built upon an ImageNet-pretrained ResNet18 [48] encoder, in which

TABLE II: Quantitative results (%) on six challenging datasets. Best results are **bold**, and the best among comparative methods are underlined. $*$ is statistical significance ($p < 0.05$) under the Wilcoxon signed-rank test [47] against the second-best model. Entries marked "-" denote results omitted for volumetric methods, which cannot be evaluated on the 2D QaTa dataset.

	Methods	LIASD [24]			ADAM [25]			BTCV [28]			BraTS [26]			TDSC [27]			QaTa [29]		
		DSC↑	BIoU↑	HD95↓	DSC↑	BIoU↑	HD95↓	DSC↑	BIoU↑	HD95↓	DSC↑	BIoU↑	HD95↓	DSC↑	BIoU↑	HD95↓	DSC↑	BIoU↑	HD95↓
CNN/ViT	U-Net [2]$_{\text{MICCAI'2015}}$	40.52	31.26	20.88	35.83	24.71	16.11	68.57	62.16	49.81	80.57	75.29	11.63	65.07	57.19	29.33	70.51	61.02	49.54
CNN/ViT	nnUNet [3]$_{\text{Nat. Methods'2021}}$	51.89	44.95	9.825	40.53	30.45	11.95	80.25	78.46	39.70	87.78	80.60	8.065	75.47	65.17	16.82	80.42	70.81	34.36
CNN/ViT	SwinUNETR [6]$_{\text{CVPR'2022}}$	43.16	32.78	19.33	38.49	27.43	13.01	77.80	75.94	40.02	89.34	82.86	8.429	72.99	66.81	18.85	-	-	-
CNN/ViT	MedNeXt [4]$_{\text{MICCAI'2023}}$	46.29	34.02	16.75	37.08	26.99	19.33	78.73	71.39	41.35	84.14	76.27	9.136	74.15	66.15	18.02	76.03	66.27	40.41
CNN/ViT	nnFormer [7]$_{\text{TIP'2023}}$	39.63	30.07	22.74	32.06	21.36	21.94	75.72	68.99	43.57	82.65	75.98	11.20	68.54	58.36	25.80	-	-	-
CNN/ViT	TransUNet [8]$_{\text{MedIA'2024}}$	42.70	32.99	21.65	34.95	25.06	17.52	75.04	71.03	43.82	81.15	76.03	13.74	61.58	53.78	30.58	72.42	60.84	49.97
SAM and Its Variants	SAM [10]$_{\text{ICCV'2023}}$	31.74	21.78	39.32	26.81	13.97	33.39	41.62	35.99	59.27	39.60	31.02	42.95	35.99	23.18	52.29	50.42	43.31	58.73
SAM and Its Variants	SAMed [17]$_{\text{arXiv'2023}}$	38.67	24.98	25.53	32.65	20.58	21.70	76.94	70.60	39.91	74.33	64.85	19.63	54.01	43.16	38.45	73.35	62.82	48.01
SAM and Its Variants	SAM-Med2D [15]$_{\text{arXiv'2023}}$	48.61	38.71	13.51	44.95	33.89	12.82	78.86	70.92	36.73	81.42	74.93	12.85	65.36	58.09	24.75	76.47	63.40	44.28
SAM and Its Variants	CPC-SAM [19]$_{\text{MICCAI'2024}}$	46.52	34.79	15.97	41.63	32.25	15.15	81.82	72.09	35.38	79.88	71.05	13.96	63.72	57.35	25.08	75.94	64.92	42.76
SAM and Its Variants	MedSAM [12]$_{\text{Nat. Commun'2024}}$	33.94	23.30	35.23	27.34	13.14	34.08	72.95	68.14	45.06	65.43	57.60	30.36	57.15	49.31	30.12	73.05	61.97	47.86
SAM and Its Variants	3DSAM-adt [14]$_{\text{MedIA'2024}}$	64.73	53.75	8.991	61.36	52.64	6.878	83.85	82.48	30.24	87.62	81.58	8.823	68.73	61.09	21.22	-	-	-
SAM and Its Variants	MA-SAM [16]$_{\text{MedIA'2024}}$	49.34	37.92	12.06	47.25	36.41	10.83	81.45	79.92	33.72	86.76	80.86	10.38	67.18	59.24	22.82	78.91	68.83	37.24
SAM and Its Variants	H-SAM [21]$_{\text{CVPR'2024}}$	43.56	32.06	18.79	34.18	24.86	18.31	77.15	70.50	42.93	78.51	74.03	14.73	59.71	53.07	26.75	74.66	63.05	45.93
SAM and Its Variants	DAPSAM [45]$_{\text{MICCAI'2024}}$	43.05	32.52	17.02	35.76	20.36	16.64	73.63	72.06	39.81	76.53	70.01	17.31	59.34	51.81	28.02	73.94	62.51	46.67
SAM and Its Variants	AutoSAMUS [22]$_{\text{MICCAI'2024}}$	45.57	33.67	16.94	39.36	29.78	14.80	78.96	72.58	36.82	81.55	75.72	13.74	61.89	54.77	25.67	75.52	63.47	44.39
SAM and Its Variants	SAM-Med3D [13]$_{\text{TNNLS'2025}}$	49.05	38.99	12.89	46.84	39.47	11.57	79.64	70.84	31.95	85.86	81.05	9.516	67.04	60.82	21.44	-	-	-
SAM and Its Variants	SAMCT [23]$_{\text{TMI'2025}}$	56.27	51.03	10.32	57.42	46.12	8.599	85.28	80.06	22.64	88.31	82.62	8.591	74.75	67.22	18.93	79.17	69.84	36.05
SAM and Its Variants	Med-SA [18]$_{\text{MedIA'2025}}$	68.62	59.11	7.752	59.43	51.37	4.328	84.54	82.22	25.87	87.10	82.06	9.033	73.57	68.03	15.38	79.54	70.15	35.68
SAM and Its Variants	SynFoC [20]$_{\text{CVPR'2025}}$	57.40	50.77	10.06	56.26	43.05	9.810	82.00	79.91	28.72	86.53	80.85	8.827	72.41	63.80	19.05	81.06	72.68	33.52
Ours	**DReAM-SAM**$^{\text{SAM (2D)}}$	**72.61***	**64.82***	**3.581***	**69.54***	**61.89***	**2.061***	**89.50***	83.82	**17.66***	94.03*	**89.21***	**5.468***	**81.35***	**72.01***	**12.24**	**84.52***	73.91	**29.41***
Ours	**DReAM-SAM**$^{\text{MedSAM (2D)}}$	71.03	63.57*	4.663*	67.95*	58.45*	2.629*	89.17*	**84.39**	18.50	**94.37***	87.33*	5.902*	80.42*	71.26	13.56	83.91*	**74.65***	30.06
Ours	**DReAM-SAM**$^{\text{SAM-Med (3D)}}$	69.92	62.24*	5.382	64.62*	56.16	3.487	86.84	83.53	19.02	90.80	86.92*	7.053	77.93*	69.38	15.07	-	-	-

TABLE III: Class-wise comparison of segmentation performance in terms of DSC (%) on the BTCV and BraTS datasets.

	Methods	BTCV [28]									BraTS [26]		
		Spl	Kid	GB	Eso	Liv	Sto	A/I/P	Pan	AG	ET	TC	WT
CNN/ViT	U-Net [2]	82.3	79.8	55.9	56.7	90.1	63.5	73.6	47.7	64.3	72.4	78.4	90.7
CNN/ViT	nnUNet [3]	90.3	89.3	68.1	77.0	93.1	80.6	82.5	72.6	68.4	84.0	87.1	92.1
CNN/ViT	SwinUNETR [6]	88.9	82.2	70.3	75.7	95.2	76.2	80.4	69.5	61.3	85.8	88.5	93.6
CNN/ViT	MedNeXt [4]	85.8	85.1	70.1	79.5	93.7	78.3	79.0	72.6	64.1	76.4	83.9	92.0
CNN/ViT	nnFormer [7]	88.5	75.9	69.8	71.2	94.2	78.0	78.9	63.7	60.8	74.6	82.7	91.5
CNN/ViT	TransUNet [8]	88.0	86.1	61.5	63.7	93.4	74.5	80.7	65.3	61.9	75.2	81.9	86.3
SAM and Its Variants	SAM [10]	40.9	50.7	29.1	30.4	51.3	42.2	52.2	42.0	35.4	35.9	38.6	44.2
SAM and Its Variants	SAMed [17]	90.9	85.4	66.4	76.0	89.9	76.5	80.2	68.4	58.3	68.5	72.4	81.9
SAM and Its Variants	SAM-Med2D [15]	90.2	87.7	76.4	80.5	89.2	72.1	79.5	70.2	65.0	75.3	80.0	88.8
SAM and Its Variants	CPC-SAM [19]	91.6	90.1	80.0	82.7	94.8	75.5	79.6	71.8	69.9	75.7	77.1	86.9
SAM and Its Variants	MedSAM [12]	90.9	82.7	63.2	73.5	85.1	70.6	75.1	65.2	49.8	60.0	63.5	72.6
SAM and Its Variants	3DSAM-adt [14]	96.7	95.0	68.2	82.1	90.1	88.2	86.1	78.6	69.1	82.0	88.6	92.1
SAM and Its Variants	MA-SAM [16]	93.3	90.9	73.4	75.3	95.7	87.1	79.3	74.6	63.0	81.8	87.2	91.3
SAM and Its Variants	H-SAM [21]	93.8	82.4	79.9	80.0	86.7	72.1	80.1	67.2	52.4	71.2	78.2	86.0
SAM and Its Variants	DAPSAM [45]	88.7	81.3	63.5	74.9	83.5	75.6	76.0	65.7	53.0	69.9	76.1	83.5
SAM and Its Variants	AutoSAMUS [22]	90.9	85.7	72.2	77.9	91.9	81.2	78.5	71.1	61.0	75.2	79.5	90.0
SAM and Its Variants	SAM-Med3D [13]	89.1	89.8	74.1	82.7	91.0	73.1	82.3	71.3	62.8	80.6	85.0	91.9
SAM and Its Variants	SAMCT [23]	95.2	92.3	79.7	83.7	93.4	85.2	86.8	79.4	71.6	84.7	88.2	91.9
SAM and Its Variants	Med-SA [18]	95.3	90.7	80.3	82.9	96.7	86.7	82.2	76.3	69.5	82.6	87.3	91.3
SAM and Its Variants	SynFoC [20]	94.1	92.4	76.5	81.9	95.3	83.6	78.2	72.7	66.0	83.1	85.9	90.5
Ours	**Ours**$^{\text{SAM (2D)}}$	98.0	**96.1**	**86.0**	85.9	**97.9**	**95.0**	87.7	**83.5**	**75.4**	**91.1**	93.2	97.8
Ours	**Ours**$^{\text{MedSAM (2D)}}$	**98.1**	95.8	85.4	**86.7**	97.5	93.4	**88.6**	83.1	73.9	90.8	**94.1**	**98.2**
Ours	**Ours**$^{\text{SAM-Med (3D)}}$	95.4	94.3	82.2	84.2	96.4	92.1	83.6	80.7	72.7	86.8	92.1	93.5

(Note) Spl: spleen, Kid: left/right kidneys, GB: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, A/I/P: aorta, inferior vena cava, and portal and splenic veins, Pan: pancreas, AG: left/right adrenal glands.

the DyRA-Block generates adaptive kernels of size $\bar{K} = 3$ from a kernel descriptor of size $K = 7$. The D^{2}NA module (Sec. III-B) is configured with a kernel size $k_d = 3$ and a maximum dilation factor $D_r = 3$ to capture multi-scale contextual information, while the temperature scaling η and embedding dimension D are set to 16 and 256, respectively. For the TAPG module (Sec. III-C), τ is set to 0.5. For the objective functions (Sec. III-D), the KD loss adopts an affinity temperature $\tau_c = 0.2$, while the RD loss employs a radius $r_d = 2$ and a boundary band $\gamma = 10$, with $\lambda = 0.5$ for $\mathcal{L}_{\text{rd}}$.

1) Comparison setup: To ensure a transparent and reproducible comparison, we implement all competing methods in our environment following the settings specified in their original publications or official protocols. In particular, each method retains the input resolution and voxel spacing adopted in its implementation, except for nnUNet (*i.e.*, the self-configuring approach). Dataset-specific modifications are limited to those required to accommodate the corresponding task, including the number of output classes and, for BraTS, the 4-channel input formed by stacking the available MRI modalities. We evaluate the model at the end of each prescribed training schedule. All results reported in this study are obtained from our reproduced experiments, and no method-specific tuning is performed beyond necessary adaptations. Furthermore, no additional pretraining on external medical datasets is performed for any method, including DReAM-SAM; each method follows the initialization prescribed by its official implementation (see Appendix Tab. XI), and subsequent task-specific training or adaptation is restricted to the training split of each dataset described in Sec. IV-A. More detailed implementation and training configurations for all compared methods are summarized in the same table.

For settings not specified in their literature, we adopt a common configuration and apply it identically to all affected methods, including DReAM-SAM. Specifically, augmentation consists of random horizontal flipping, random rotation, and random intensity shift and scaling, each with probability 0.5. In addition, target-present and target-absent samples are drawn at a 1:1 ratio for all methods to mitigate severe foreground–background imbalance and stabilize training. For prompted methods, the prompt type listed in Appendix Tab. XI is used identically during training and inference, with point prompts placed at the center of the GT region and box prompts given as the tight GT bounding box.

The 3D methods, namely SwinUNETR, 3DSAM-adt, nnFormer, and SAM-Med3D, operate on volumetric inputs and

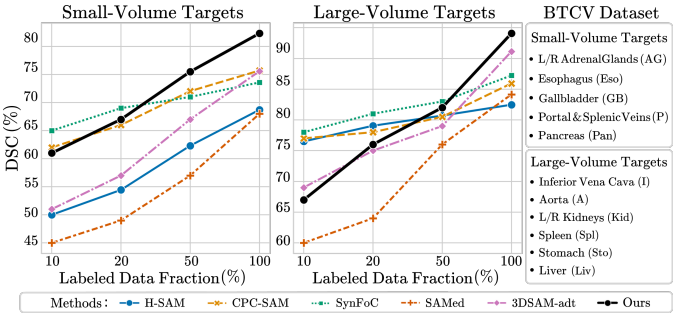

Fig. 6: Segmentation performance (DSC, %) on BTCV dataset under varying labeled data fractions, evaluated separately on small- and large-volume target organs to assess data efficiency.

therefore exploit inter-slice context, which is unavailable in QaTa-COV19 (2D chest radiographs). As adapting them with architectural modifications would depart from the published methods, we exclude them—along with our own DReAM-SAM$^{\text{SAM-Med (3D)}}$—and mark these entries as "-" in Tab. II.

2) Evaluation metrics: To comprehensively assess segmentation performance, we adopt the Dice Similarity Coefficient (DSC), boundary IoU (BIoU) [49], and the 95% Hausdorff Distance (HD95). DSC and BIoU are set to 1 when the predicted and GT masks are both empty, and to 0 when only one of them is empty. HD95 is averaged only over non-empty mask pairs, as the distance is otherwise undefined. We also report the false-positive (FP) rate, defined as the proportion of non-empty predictions among target-absent slices or volumes.

All metrics are computed on 3D volumes reassembled at the native spacing of each scan, except the slice-level FP rate, which is computed on individual slices. QaTa is evaluated without reassembly, its data being inherently 2D, and values on the multi-class BTCV and BraTS are macro-averaged over target classes. In addition, the Wilcoxon signed-rank test [47] with $p < 0.05$ is conducted to verify statistical significance.

D. Comparison with State-of-the-arts

For the quantitative assessment, we compare the DReAM-SAM against two major groups of state-of-the-art (SOTA) methods: (i) CNN/ViT-based models [2]–[4], [6]–[8] and (ii) SAM [10] and its variants [12]–[23], [45].

Fig. 5 presents a qualitative comparison of various methods. DReAM-SAM produces tighter and more coherent boundaries on small or low-contrast targets, where competing methods often miss the lesion entirely or over-segment into surrounding tissue. Tab. II reports the quantitative comparison results. Overall, DReAM-SAM consistently achieves the best DSC, BIoU, and HD95 across all benchmarks. It is noteworthy that the DSC margin over the second-best method is largest on datasets dominated by small, ill-defined targets (+8.18% on ADAM, +5.88% on TDSC, and +3.99% on LIASD), whereas it is comparatively narrower on QaTa (+3.46%), whose pneumonia regions span a much larger spatial extent. In Tab. III, we further provide class-specific DSC on BTCV and BraTS. On BTCV, DReAM-SAM attains top performance not only for large organs such as the Liv (97.9%) and Kid (96.1%), but also yields clear gains on smaller structures, improving GB (+5.7%) and AG (+3.8%) segmentation over the second-best method. For BraTS, DReAM-SAM delivers clear advantages, with DSC performance of +5.3% (ET), +4.6% (TC), and +4.2% (WT) compared to the second-best method, with the largest gain occurring on the ET—the smallest and most heterogeneous sub-region—further supporting that the framework's benefit grows with target difficulty.

TABLE IV: Comparison of computational cost and DSC on LIASD ($N = 1$) and BTCV ($N = 13$) datasets, reported as LIASD/BTCV. "#Total" and "#Train" denote the total and trainable parameters, "FLOPs" denotes floating-point operations, "Mem." denotes GPU memory usage during inference, and "Time" denotes inference time.

Methods	#Total (M)	#Train (M)	FLOPs (G)	Mem. (GB)	Time (ms)	DSC (%)
nnUNet	22.6/22.6	22.6/22.6	773.2/774.1	0.75/0.75	36.4/36.6	51.9/80.3
SwinUNETR	62.2/62.2	62.2/62.2	1467.1/1469.6	4.37/4.61	360.3/362.2	43.2/77.8
TransUNet	94.7/94.7	94.7/94.7	1553.5/1557.1	2.31/2.31	216.5/217.9	42.7/75.0
SAMed	93.9/93.9	4.2/4.2	747.5/790.6	2.70/3.06	154.9/226.6	38.7/76.9
H-SAM	111.2/116.1	21.5/26.4	753.1/758.2	2.79/2.80	155.1/157.3	43.6/77.2
AutoSAMUS	139.9/139.8	48.5/48.5	1200.7/1514.6	2.91/3.43	233.2/373.7	45.6/79.0
SAMCT	125.7/125.7	35.2/35.2	2018.9/2722.1	5.01/5.20	611.1/927.9	56.3/85.3
Med-SA	719.9/719.9	82.9/82.9	808.2/851.5	2.75/2.75	163.1/242.5	68.6/84.5
SynFoC	93.7/95.4	4.1/5.8	743.9/743.9	2.70/2.71	149.4/150.7	57.4/82.0
Ours	156.8/171.0	67.1/81.3	1366.6/1522.7	3.09/3.48	164.6/353.8	72.6/89.5

(Note) All methods are evaluated with 1024×1024 input, except the 3D SwinUNETR.

1) Size-based evaluation under limited supervision: Fig. 6 exhibits DSC scores across varying labeled data fractions (10%-100% labeled data) on BTCV, where target organs are grouped into small-volume (SV) (mean: 13,263 mm^3) and large-volume (LV) (mean: 198,267 mm^3), a $\approx 15\times$ size gap. We compare DReAM-SAM against two approach groups: adaptation methods (ADM: SAMed, 3DSAM-adt), which are trained only on the limited labeled data, and semi-supervised methods (SSM: H-SAM, CPC-SAM, SynFoC) that are trained on both labeled and remaining unlabeled data. In Fig. 6, SSMs show stable DSC performance on LV targets, maintaining consistent accuracy across all label fractions. However, their performance drops sharply for SV targets, indicating that even semi-supervised schemes struggle with small or ambiguous boundaries despite using additive unlabeled data. In contrast, DReAM-SAM outperforms both ADMs and SSMs on SV targets once the labeled fraction exceeds 50%, achieving the best accuracy among all comparative methods. This is particularly notable given that DReAM-SAM does not employ any semi-supervised strategy, yet still verifies superior robustness to both scarce annotations and small target volumes.

2) Computational efficiency: We further analyze the trade-off between segmentation performance and computational cost on LIASD and BTCV in Tab. IV. Among the CNN-augmented methods, DReAM-SAM achieves higher DSC with lower inference latency than both AutoSAMUS and SAMCT, while maintaining computational costs comparable to AutoSAMUS and substantially lower than SAMCT. In particular, DReAM-SAM improves DSC over AutoSAMUS by 27.0/10.5 percentage points while reducing inference time from 233.2/373.7 ms to 164.6/353.8 ms. Although nnUNet is considerably more lightweight, DReAM-SAM improves DSC by 20.7 and 9.2 percentage points on LIASD and BTCV, respectively. It also

TABLE V: Ablation study of DReAM-SAM components in terms of DSC (%) across five datasets.

(a) Ablation of Each Module under Our Self-Prompting Mechanism (DReAM-SAM)

Cases	TAPG	TEC	D^2NA	DyRA	$\mathcal{L}_{kd}$	$\mathcal{L}_{rd}$	ADAM	BTCV	BraTS	TDSC	QaTa
1	✓	-	-	-	-	-	35.11	76.23	73.46	56.49	73.41
2	✓	✓	-	-	-	-	48.72	79.36	82.14	68.34	74.02
3	✓	✓	✓	-	-	-	60.31	83.25	87.32	73.67	78.50
4	✓	✓	-	✓	-	-	62.14	84.02	87.91	74.35	78.91
5	✓	✓	-	✓	✓	-	64.05	87.00	89.95	77.06	80.73
6	✓	✓	-	✓	-	✓	65.37	87.03	90.82	77.28	81.10
7	✓	✓	-	✓	✓	✓	66.35	88.26	92.04	79.37	82.36
Ours	✓	✓	✓	✓	✓	✓	**69.54**	**89.50**	**94.03**	**81.35**	**84.52**

(b) Evaluation under Manual Prompting with a Single GT Point (SAM)

Cases	GT	FT	D^2NA	DyRA	$\mathcal{L}_{kd}$	$\mathcal{L}_{rd}$	ADAM	BTCV	BraTS	TDSC	QaTa
1	✓	-	-	-	-	-	26.81	41.62	39.60	35.99	50.42
2	✓	✓	-	-	-	-	50.24	82.16	85.33	70.57	67.21
3	✓	✓	✓	-	-	-	65.13	85.17	87.94	75.16	81.03
4	✓	✓	-	✓	-	-	66.59	85.46	88.20	77.01	82.67
5	✓	✓	-	✓	✓	-	68.43	88.91	90.02	78.34	83.19
6	✓	✓	-	✓	-	✓	69.26	89.25	90.37	80.71	84.22
7	✓	✓	-	✓	✓	✓	71.84	90.58	91.26	81.29	85.03
8	✓	✓	✓	✓	✓	✓	**73.18**	**92.64**	**94.30**	**83.67**	**86.95**

(**Note**) FT: Fine-tuning SAM with the single GT point used as a manual prompt.

TABLE VI: Component-level ablation of core design choices in terms of DSC (%). Each block replaces a single design element, keeping all others fixed. "#Params" is the number of trainable parameters (M) of the corresponding configuration, and "ΔParams" denotes its difference from our full model.

Cases	ADAM	BTCV	BraTS	TDSC	QaTa	#Params	ΔParams
(a) DyRA-Block: kernel composition							
1. Static kernel (*i.e.*, plain CNNs)	60.31	83.25	87.32	73.67	78.50	58.816M	–8.309M
2. Static kernel (param-matched)	61.52	84.13	88.20	74.58	79.26	67.117M	–0.008M
3. Shared kernel only	64.85	85.44	90.02	78.46	82.31	60.497M	–6.628M
4. Target-specific kernel only	66.24	86.15	91.28	79.05	83.20	63.981M	–3.143M
5. Shared + specific (w/o gate)	67.96	87.31	92.18	79.88	83.52	67.123M	–0.001M
6. Ours w/o $\mathcal{L}_{kd}$	68.40	87.65	92.51	80.42	83.91	67.124M	±0M
Shared + gated specific (Ours)	**69.54**	**89.50**	**94.03**	**81.35**	**84.52**	67.124M	–
(b) D^2NA: interaction streams							
1. Concatenation (w/o D^2NA)	61.36	82.59	87.14	75.33	77.62	41.637M	–25.487M
2. Self-DNA only	65.21	86.94	90.61	79.05	80.94	60.272M	–6.853M
3. Cross-DNA only	67.33	88.17	91.25	80.61	83.68	60.272M	–6.853M
4. Uniform dilation weighting	67.71	88.06	92.44	80.79	82.92	67.112M	–0.012M
Self + Cross, query-adaptive (Ours)	**69.54**	**89.50**	**94.03**	**81.35**	**84.52**	67.124M	–
(c) Self-prompting: prompt composition							
1. Dense prompt only	67.42	87.13	92.58	79.86	83.14	66.076M	–1.049M
2. Sparse prompt only	66.87	84.42	89.75	77.51	81.01	67.119M	–0.005M
3. Dense + Sparse (w/o gating)	67.55	88.29	93.08	80.53	82.79	67.124M	±0M
Dense + Sparse, gated (Ours)	**69.54**	**89.50**	**94.03**	**81.35**	**84.52**	67.124M	–

outperforms Med-SA by 4.0/5.0 percentage points while using approximately 4.2–4.6× fewer total parameters. Overall, these results indicate a favorable accuracy–efficiency trade-off and suggest that the performance gains are not simply attributable to increased model capacity or computation.

Meanwhile, all efficiency measurements are obtained on an NVIDIA Quadro RTX 8000. Therefore, while the results demonstrate feasible GPU-based inference under the evaluated workstation setting, they do not directly establish real-time or resource-constrained clinical deployment, which requires further device-specific evaluation and optimization.

E. Ablation Study

1) Effectiveness of core modules: Tab. V presents the ablation studies on the key components of DReAM-SAM.

TABLE VII: Contribution of the SAM branch to the overall framework. DSC (%) is reported on five datasets, and "#Params" denotes the number of trainable parameters (M).

Settings	ACNN Branch	SAM Encoder	D^2NA Module	Self-Prompt	Mask Decoder	$\mathcal{L}_{total}$	ADAM	BTCV	BraTS	TDSC	QaTa	#Params (Learnable)
ACNN Only	✓	✗	✗	✓	✓	✓	63.37	81.24	87.22	76.28	77.89	32.234M
DReAM-SAM	✓	✓	✓	✓	✓	✓	**69.54**	**89.50**	**94.03**	**81.35**	**84.52**	67.124M

In (a), we investigate the contribution of each module under our self-prompting setting. Adding TEC to a TAPG-only baseline (Case 1→2) yields the largest gain on the tiny-target ADAM (+13.61%), where target-existence priors are most informative, but only a modest improvement on multi-class BTCV (+3.13%) since class disambiguation alone cannot resolve anatomically similar organs. Building on Case 2, both D^2NA (Case 3) and DyRA (Case 4) yield substantial gains (+11.59% and +13.42% on ADAM). D^2NA primarily benefits low-contrast targets (+5.33% on TDSC) by harmonizing SAM's global semantics with ACNN's localized cues, whereas DyRA shows stronger advantages on morphologically diverse targets (+5.77% on BraTS), where its target-conditioned kernels better adapt to varying shape and scale. On top of DyRA, $\mathcal{L}_{kd}$ (Case 5) and $\mathcal{L}_{rd}$ (Case 6) provide complementary regularization: $\mathcal{L}_{kd}$ yields the largest gains on multi-target datasets with overlapping classes (+2.98% on BTCV, +2.04% on BraTS over Case 4) by preventing target-specific kernels from collapsing, while $\mathcal{L}_{rd}$ contributes most on datasets with ambiguous boundaries (+2.93% on TDSC, +2.19% on QaTa) via boundary-aware mask supervision. Their combination (Case 7) yields further additive gains, and the full model (Ours) attains the best DSC across all five datasets.

In (b), we further assess the same modules under manual prompting, where the base SAM receives a single GT point. Even in this strong setup, our modules yield dramatic gains (Case 2→8: +22.94% on ADAM, +19.74% on QaTa), indicating that the improvements stem from a genuine enhancement of SAM's representational capacity rather than from the self-prompting regime alone. Notably, **(a) Ours** recovers most of the GT-prompted **(b) Case 8** performance without any manual prompts (within 0.3% on BraTS and 2–4% on the remaining datasets), suggesting that TAPG's self-derived localization signals largely substitute for a manually placed GT point.

2) Analysis of module-level design choices: In Tab. VI, we further investigate the individual design choices within DyRA, D^2NA, and self-prompting by modifying one design element at a time while keeping all remaining components fixed.

(a) DyRA-Block composition. We first examine whether the benefit of DyRA is primarily attributable to its target-conditioned kernel formulation rather than to increased convolutional capacity. In Case 1, the DyRA-Block is removed entirely, leaving the original ResNet18 encoder with standard static convolutions. In Case 2, the encoder remains static, but the corresponding blocks are widened with two 3×3 convolutional layers to approximately match the parameter count of the DReAM-SAM. Increasing the static encoder from 58.816M to 67.117M parameters yields only modest DSC gains, and the parameter-matched variant still substantially

TABLE VIII: TEC accuracy (%) under module ablations.

Cases	Settings	LIASD	ADAM	BTCV	BraTS	TDSC
1	Base (ResNet18)	90.37	89.35	85.92	81.26	86.27
2	TEC w/o DyRA and D^{2}NA	89.24	86.34	85.66	79.21	85.99
3	TEC w/o DyRA	95.89	93.52	87.48	83.88	87.64
4	TEC w/o D^{2}NA	96.07	92.61	86.49	82.56	87.04
Ours	TEC trained within DReAM-SAM	**97.83**	**97.02**	**93.55**	**89.71**	**94.82**

TABLE IX: FP rate (%) on target-absent slices and volumes.

Existence gating	Slice-level					Volume-level	
	LIASD	ADAM	BTCV	BraTS	TDSC	LIASD	ADAM
GT (oracle TEC)	0.02	0.23	0.35	0.19	0.38	2.00	16.67
None (no gating)	64.12	80.71	58.80	66.79	90.67	92.00	100.00
TEC (ours)	**1.91**	**2.28**	**6.35**	**10.30**	**6.12**	**18.00**	**33.33**

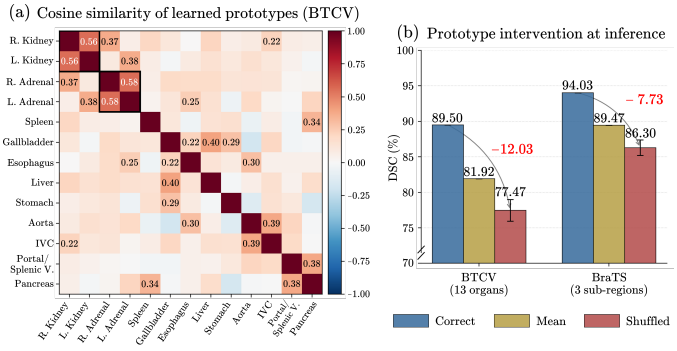

Fig. 7: Analysis of the learned target prototypes $\{\mathbf{c}_n\}$: (a) Pairwise cosine similarities on BTCV; black boxes mark the two bilateral organ pairs, and values below 0.22 are omitted for readability. (b) Prototype intervention at inference without retraining. "Mean" replaces every $\mathbf{c}_n$ with the mean prototype, while "Shuffled" randomly permutes them across targets.

underperforms DyRA despite having nearly identical capacity (67.117M vs. 67.124M). Relative to Case 2, DyRA improves DSC by an average of 6.25% across all datasets, with gains ranging from +5.26% to +8.02%, indicating that the gains arise from the target-conditioned kernel formulation rather than additional convolutional capacity.

Using only the shared kernel (Case 3) or only the target-specific kernel (Case 4) also consistently underperforms the full formulation, suggesting that neither component alone is sufficient. Instead, combining shared cross-target patterns with target-dependent filtering provides a more effective representation. Removing the adaptive gate (Case 5) further reduces DSC by 1.00–2.19% across datasets, supporting the benefit of dynamically balancing the shared and specific components. Finally, ablating $\mathcal{L}_{\text{kd}}$ (Case 6) causes consistent performance degradation, particularly on BTCV (–1.85%) and BraTS (–1.52%), highlighting its role in preserving the distinguishability of target-specific kernel components.

(b) D^{2}NA interaction design. We evaluate the necessity of D^{2}NA by replacing the module with channel-wise concatenation (Case 1). This baseline yields the lowest DSC across all datasets, trailing the full D^{2}NA formulation by 6.02–8.18%, suggesting that simple feature aggregation is insufficient to effectively integrate representations from the ACNN and SAM branches. Introducing either self-DNA (Case 2) or cross-DNA (Case 3) leads to substantial improvements. Among the two single-stream variants, cross-DNA consistently outperforms self-DNA across all datasets, *e.g.*, 67.33% vs. 65.21% on ADAM and 88.17% vs. 86.94% on BTCV, highlighting the importance of explicit inter-branch information exchange. Combining self-DNA and cross-DNA yields additional gains, outperforming Case 3 by 0.74–2.78% and achieving the best results across all datasets. These findings indicate that intra-branch refinement and inter-branch interaction play complementary roles. We further examine the query-adaptive dilation mechanism by replacing the learned dilation weights with uniform weighting (Case 4). Despite nearly identical parameter counts (67.112M vs. 67.124M), uniform weighting results in a 0.56–1.83% decrease in DSC, supporting the effectiveness of query-adaptive multi-scale contextual aggregation.

(c) Self-prompt composition. We further investigate the contributions of the individual prompt components within the self-prompting mechanism. Dense prompting alone (Case 1) consistently outperforms sparse prompting alone (Case 2), with notable margins on BTCV (87.13% vs. 84.42%) and BraTS (92.58% vs. 89.75%). This advantage is consistent with dense prompts leveraging the spatially rich, target-conditioned representations produced by DyRA, whereas sparse prompts compress such information into a limited set of point-level cues. Combining these prompts without target-existence gating (Case 3) improves over the sparse-only variant on all datasets and over the dense-only variant on four of the five datasets, indicating that the two prompt types provide complementary guidance. Incorporating TEC-derived target-existence gating yields additional improvements across all datasets, with an average DSC gain of 1.34% over the ungated formulation (range: 0.82–1.99%). These results highlight the importance of target-existence gating for robust automatic prompting by suppressing spurious guidance when the target is absent.

3) Contribution of the SAM branch: Beyond the component-wise ablations, we isolate the contribution of the SAM branch by removing the SAM encoder while preserving the remainder of the DReAM-SAM framework. Since D^{2}NA explicitly operates between the ACNN and SAM representations, it is necessarily omitted in this setting. As reported in Tab. VII, incorporating the SAM branch consistently improves DSC by 5.07–8.26% across all five datasets, increasing the average DSC from 77.20% to 83.79%. Notably, the ACNN-only variant reaches this level with fewer than half the trainable parameters of the full model (32.234M vs. 67.124M), indicating that the ACNN branch itself learns effective target-adaptive representations independently of the SAM encoder. The gain from the SAM branch is largest on the multi-class BTCV (+8.26%) and smallest on the locally defined TDSC (+5.07%), suggesting that its contribution depends on the contextual requirements of each task.

4) Fidelity of TEC guidance: Tab. VIII reports the TEC classification accuracy under various ablation settings, with ResNet18 trained independently on each dataset serving as a reference (Case 1). Interestingly, jointly training TEC with TAPG but without DyRA and D^{2}NA (Case 2) drops accuracy

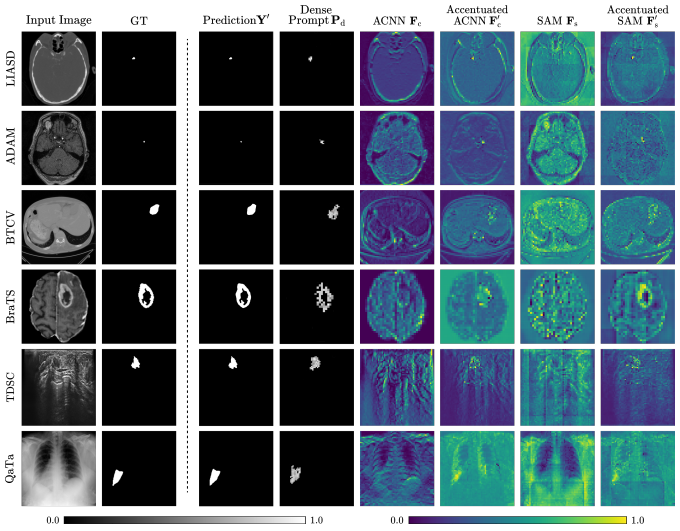

Fig. 8: Visualization of feature responses from both ACNN and SAM branches before and after accentuation, along with generated dense prompts and final predictions.

below this baseline (-1.34% on average), suggesting that the vanilla CNN encoder alone offers limited representational support for the classification objective. Reintroducing D^2NA (Case 3) or DyRA (Case 4) reverses this trend, with average gains of +4.39% and +3.66% over Case 2, respectively. The slight edge of D^2NA suggests that harmonizing SAM's global semantics with ACNN's localized cues benefits slice-level existence prediction, whereas DyRA mainly strengthens target-conditioned local discrimination. Consequently, the full DReAM-SAM achieves the highest accuracy across all datasets (+8% over the standalone baseline). Beyond classification accuracy, Appendix Tab. XII provides a more comprehensive assessment of TEC reliability. These results further characterize TEC discrimination and calibration under class imbalance and support its use as a semantic prior for TAPG.

5) Effect of existence gating on false positives: We additionally examine the effect of the TEC-based gating mechanism on FP predictions. Tab. IX reports the FP rate on target-absent slices and volumes under three gating configurations: no gating, TEC-based gating, and oracle gating using GT existence labels. Without gating, the FP rate reaches 58.80–90.67% at the slice level, whereas TEC substantially reduces the FP rate by 33.6× on LIASD and 35.4× on ADAM, both of which contain small and low-contrast targets. The FP rates under TEC are comparable to those obtained with GT gating, with a difference of 1.89–10.11% at the slice level. A similar trend is observed at the volume level, demonstrating that TEC effectively suppresses FP predictions. The lesion-level analysis in Appendix Tab. XIII further demonstrates that this FP reduction is achieved while maintaining high lesion recall.

F. Analysis of Learned Target Prototypes

To examine the information captured by the learned target prototypes, we analyze their pairwise similarities and perform controlled prototype interventions at inference. On BTCV, anatomically corresponding bilateral organs exhibit among the highest off-diagonal prototype similarities (see Fig. 7(a)), including the kidneys (0.56) and adrenal glands (0.58), suggesting that the learned prototype space reflects meaningful similarities among anatomically related targets.

In Fig. 7(b), we further evaluate the functional role of the prototypes by replacing them at inference without retraining. Specifically, replacing every target prototype with the mean prototype reduces DSC from 89.50% to 81.92% on BTCV and from 94.03% to 89.47% on BraTS, whereas randomly shuffling the prototypes across targets further decreases DSC to 77.47% and 86.30%, respectively. These consistent degradations indicate that a single shared prototype cannot fully represent the target-dependent information encoded by $\mathbf{c}_n$, while the larger degradation under prototype shuffling highlights the importance of preserving the correct target–prototype correspondence for generating effective target-conditioned kernels.

TABLE X: Cross-institution and cross-modality DSC (%). For cross-institution evaluation, models are trained on the six-institution internal training set of LIASD and evaluated on two completely independent external institutions (Ext. A and B); for cross-modality evaluation, they are trained on one dataset and evaluated on the other (CTA ↔ ToF-MRA). No target-domain adaptation is applied in either setting. "–": transfer is undefined, as SAM and MedSAM are evaluated with their released weights without task-specific training.

Methods	Cross-Institution (LIASD)			Cross-Modality (LIASD ↔ ADAM)	
	Internal	Ext. A	Ext. B	LIASD→ADAM	ADAM→LIASD
nnUNet	51.89	69.07	46.33	19.67	25.36
SAM	31.74	23.75	19.51	–	–
MedSAM	33.94	28.94	23.45	–	–
SAMed	38.67	58.13	33.52	31.22	38.58
Med-SA	68.62	71.34	51.97	32.49	42.83
DReAM-SAM	**72.61**	**76.49**	**58.22**	**45.25**	**52.16**

G. Feature Visualization and Analysis

To better understand the feature enhancement in DReAM-SAM, we visualize the averaged intermediate feature maps from both ACNN ($\mathbf{F}_c$, $\mathbf{F}'_c$) and SAM ($\mathbf{F}_s$, $\mathbf{F}'_s$) branches, along with the surrogate prompt $\mathbf{P}_d$ and the final prediction $\mathbf{Y}'$. As shown in Fig. 8, it is noteworthy that the generated prompt $\mathbf{P}_d$ closely resembles the GT in both location and shape, supporting its role in precise prompt guidance. While $\mathbf{F}_c$ and $\mathbf{F}_s$ mainly capture local textures and global semantics, their accentuated counterparts ($\mathbf{F}'_c$ and $\mathbf{F}'_s$) exhibit fine appearances as well as a pronounced focus on target locations. These observations suggest our modules collaboratively refine complementary representations, thereby strengthening target discrimination and sensitivity to small or ambiguous structures.

V. Discussion

A. Generalization Beyond the Training Distribution

DReAM-SAM is designed to adapt a segmentation foundation model for target-aware, prompt-free inference, where target-aware guidance is automatically derived from the input image based on learned target priors, rather than provided via user prompts. This naturally raises the question of *whether*

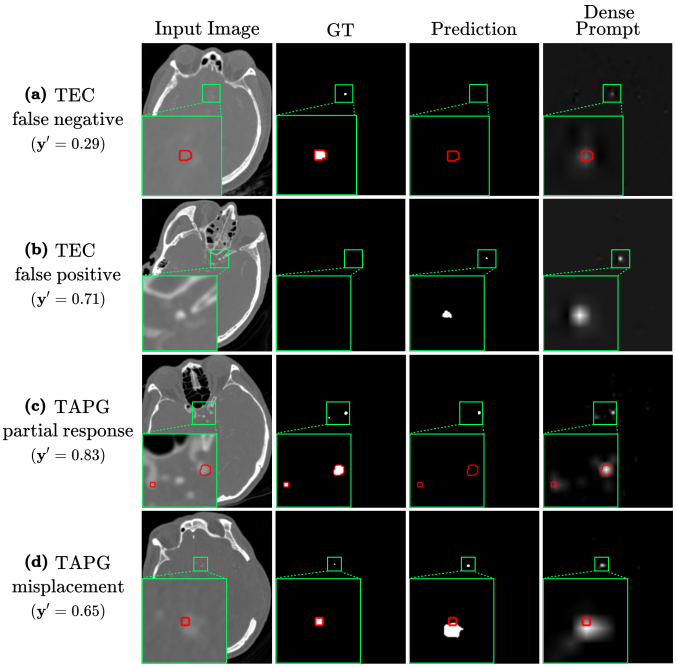

Fig. 9: Representative failure cases on LIASD, annotated with the existence probability $\mathbf{y}'$ predicted by TEC. Red contours are the GT boundary, and green boxes are the zoomed regions.

the observed gains reflect robust adaptation of the pretrained foundation-model representations or primarily dataset-specific fitting. To investigate this, we evaluate DReAM-SAM under two complementary distribution shifts in aneurysm segmentation: (i) cross-institution evaluation between the internal and external LIASD cohorts, and (ii) cross-modality transfer between LIASD (CTA) and ADAM (ToF-MRA). The results are summarized in Tab. X.

For cross-institution evaluation, methods requiring task-specific training or adaptation are trained on the LIASD internal cohort collected from six institutions and evaluated on the corresponding internal test set and two independent external cohorts [24], *i.e.*, External A (Ext. A) and External B (Ext. B), without further adaptation. The two external cohorts comprise 71 (24 normal) and 67 (22 normal) CTA cases acquired at two additional institutions, using scanner models and a slice thickness of 0.9 mm that do not appear in the internal cohort. On these cohorts, we compare DReAM-SAM with nnUNet, SAM, MedSAM, SAMed, and Med-SA, representing conventional supervised, foundation-model, and SAM-adaptation baselines. DReAM-SAM achieves the highest DSC scores across all three cohorts, with 72.61%, 76.49%, and 58.22% on the internal, Ext. A, and Ext. B sets, respectively. For cross-modality evaluation, we perform direct transfer between LIASD CTA and ADAM ToF-MRA in both directions without target-domain fine-tuning or adaptation. Under this more substantial distribution shift, DReAM-SAM achieves the best performance, reaching 45.25% DSC for LIASD→ADAM and 52.16% for ADAM→LIASD, outperforming the strongest baseline, Med-SA, by 12.76 and 9.33 percentage points, respectively.

Overall, these experiments complement the original six-dataset evaluation: while the latter demonstrates the broad applicability of a common adaptation strategy across heterogeneous segmentation tasks, the additional experiments assess robustness under explicit institutional and modality shifts. Importantly, DReAM-SAM remains a supervised closed-set adaptation framework; thus, these results should not be interpreted as evidence of open-vocabulary or arbitrary unseen-target segmentation. Rather, they show that, for a fixed clinical target, the proposed adaptation strategy maintains a consistent advantage beyond the training distribution. Generalization to previously unseen semantic targets and broader external cohorts remains an important direction for future work.

B. Failure Case Analysis

Despite the overall effectiveness of our DReAM-SAM, its performance can still be limited by errors in target-existence estimation and prompt localization, as illustrated by the representative failure cases on LIASD in Fig. 9. In Fig. 9(a), TEC produces a false-negative existence prediction ($\mathbf{y}' = 0.29$), resulting in insufficient target guidance and a missed lesion. Conversely, the false-positive TEC response in Fig. 9(b) ($\mathbf{y}' = 0.71$) produces a localized activation in the dense prompt, leading to a spurious foreground prediction. These cases indicate that errors in TEC can affect whether target-related evidence is appropriately activated.

Figs. 9(c) and (d) illustrate complementary failure modes of TAPG. In the multi-lesion case, TAPG responds strongly to the dominant lesion but only weakly to the smaller lesion, resulting in partial detection. In Fig. 9(d), spatial misalignment of the dense prompt leads to a shifted and enlarged segmentation. These observations highlight the distinct roles of the two components: TEC determines whether target evidence is activated, whereas TAPG determines where that evidence is spatially concentrated. Future work could therefore improve robustness to small and multifocal lesions via uncertainty-aware existence estimation and multi-scale or multi-peak prompt generation.

VI. Conclusion

In this work, we proposed DReAM-SAM, a fully automatic SAM adaptation framework for fine-grained MIS that introduces target- and region-adaptive representation learning. DReAM-SAM adapts local feature extraction to target identity and morphology, integrates complementary ACNN and SAM representations through query-adaptive multi-scale interaction, and generates surrogate prompts from inferred target-aware cues without manual intervention. Extensive experiments across six MIS datasets and diverse medical imaging modalities demonstrated consistent improvements over SOTA SAM-based and task-specific segmentation methods, with particularly strong performance on small, heterogeneous, and ambiguous targets. Additional cross-institution and cross-modality evaluations further demonstrated the robustness of the proposed adaptation strategy under explicit distribution shifts. These results support the effectiveness of explicitly incorporating target identity and regional anatomical context into SAM adaptation for challenging MIS.

APPENDIX

TABLE XI: Detailed configurations of all compared methods.

Method	Backbone	Init.	Input	Prompt	Opt./LR/B/Schedule
U-Net	CNN	✗	572×572	–	SGD/1e-3/16/300
nnUNet	CNN	✗	(self-config.)	–	SGD/1e-2/–/1000
SwinUNETR	Swin Transformer	✗	128×128×128	–	AdamW/8e-4/4/800
MedNeXt	2D ConvNeXt	✗	512×512	–	AdamW/1e-3/12/1000
nnFormer	Swin Transformer+CNN	✗	128×128×64	–	SGD/1e-2/2/1000
TransUNet	ViT-B + ResNet50	✓	224×224	–	SGD/1e-2/24/150
SAM	SAM (ViT-B)	✓	1024×1024	Point	–
SAMed	SAM (ViT-B)	✓	512×512	–	AdamW/5e-3/12/200
SAM-Med2D	SAM (ViT-B)	✓	256×256	Point	Adam/1e-4/50/12
CPC-SAM	SAM (ViT-B)	✓	512×512	–	AdamW/1e-3/12/10k
MedSAM	MedSAM (ViT-B)	✓	1024×1024	Box	–
3DSAM-adt	SAM (ViT-B)	✓	128×128×128	Point	AdamW/1e-4/1/200
MA-SAM	SAM (ViT-H)	✓	512×512	–	Adam/8e-4/24/400
H-SAM	SAM (ViT-B)	✓	224×224	–	AdamW/2.5e-3/16/300
DAPSAM	SAM (ViT-B)	✓	384×384	Auto	AdamW/5e-4/8/200
AutoSAMUS	SAM (ViT-B)+CNN	✓	256×256	Auto	Adam/5e-4/8/200
SAM-Med3D	SAM-Med3D	✓	128×128×128	Point	AdamW/8e-5/12/200
SAMCT	SAM (ViT-B)+CNN	✓	256×256	Auto	Adam/6e-4/48/50
Med-SA	SAM (ViT-H)	✓	1024×1024	Point	Adam/1e-4/32/60
SynFoC	MedSAM (ViT-B)	✓	512×512	–	AdamW/1e-4/8/60k iter.
Ours$^{\text{SAM (2D)}}$	SAM (ViT-B) + ResNet18	✓	1024×1024	Auto	AdamW/1e-5/4/50
Ours$^{\text{MedSAM (2D)}}$	MedSAM (ViT-B) + ResNet18	✓	1024×1024	Auto	AdamW/1e-5/4/50
Ours$^{\text{SAM-Med (3D)}}$	SAM-Med3D (ViT-B) + ResNet18	✓	128×128×128	Auto	AdamW/1e-5/4/50

(**Note**) Init. (✓: pretrained, ✗: trained from scratch); Prompt (Point/Box: GT-derived, Auto: self-generated, "–": none); Opt.: optimizer, LR: learning rate, B: batch size

TABLE XII: Detailed TEC performance (%) across datasets. Prevalence denotes the percentage of target-present slices in the test set. Expected calibration error (ECE) measures the discrepancy between TEC-predicted probabilities $\mathbf{y}'$ and GT labels $\mathbf{y}$, computed over 10 equal-width bins.

Metric	LIASD	ADAM	BTCV	BraTS	TDSC
Prevalence	2.68	5.93	27.07	32.38	10.81
Sensitivity (↑)	81.49	80.52	90.69	86.25	91.76
Specificity (↑)	98.28	98.06	95.59	93.26	95.19
AUROC (↑)	93.62	92.32	96.58	92.77	96.48
AUPRC (↑)	76.21	70.84	87.31	83.02	90.58
ECE (↓)	4.18	4.92	2.17	3.54	2.63

TABLE XIII: Lesion-level detection on LIASD and ADAM. Recall is stratified by lesion size and count per scan, with the number of GT lesions in each stratum given in parentheses. FP/scan is averaged over all test scans, including normal cases.

Dataset	Recall by size			Recall by multiplicity		Overall	FP
	≤3 mm	3–7 mm	>7 mm	single	multiple	recall↑	/scan↓
LIASD	75.00 (4)	86.96 (46)	98.68 (76)	95.29 (85)	90.24 (41)	93.65 (126)	0.08
ADAM	76.47 (17)	93.75 (16)	100.00 (5)	90.00 (20)	83.33 (18)	86.84 (38)	0.15

REFERENCES

[1] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10 076–10 095, 2024.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[3] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[4] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, and K. H. Maier-Hein, "Mednext: transformer-driven scaling of convnets for medical image segmentation," in *MICCAI*. Springer, 2023, pp. 405–415.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[6] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3D medical image analysis," in *CVPR*, 2022, pp. 20 730–20 740.

[7] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, "nnFormer: Volumetric medical image segmentation via a 3D transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 4036–4045, 2023.

[8] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, "TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers," *Med. Image Anal.*, p. 103280, 2024.

[9] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[11] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: an experimental study," *Med. Image Anal.*, vol. 89, p. 102918, 2023.

[12] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nat. Commun.*, vol. 15, no. 1, p. 654, 2024.

[13] H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, J. Chen, Y. Su, Z. Huang, Y. Shen, B. Fu, S. Zhang, and J. He, "SAM-Med3D: A vision foundation model for general-purpose segmentation on volumetric medical images," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 36, no. 10, pp. 17 599–17 612, 2025.

[14] S. Gong, Y. Zhong, W. Ma, J. Li, Z. Wang, J. Zhang *et al.*, "3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation," *Med. Image Anal.*, vol. 98, p. 103324, 2024.

[15] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang *et al.*, "SAM-Med2D," *arXiv preprint arXiv:2308.16184*, 2023.

[16] C. Chen, J. Miao, D. Wu, A. Zhong, Z. Yan, S. Kim, J. Hu, Z. Liu *et al.*, "MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation," *Med. Image Anal.*, vol. 98, p. 103310, 2024.

[17] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.13785*, 2023.

[18] J. Wu, Z. Wang, M. Hong, W. Ji, H. Fu, Y. Xu, M. Xu, and Y. Jin, "Medical SAM adapter: Adapting segment anything model for medical image segmentation," *Med. Image Anal.*, vol. 102, p. 103547, 2025.

[19] J. Miao, C. Chen, K. Zhang, J. Chuai *et al.*, "Cross prompting consistency with segment anything model for semi-supervised medical image segmentation," in *MICCAI*. Springer, 2024, pp. 167–177.

[20] Q. Ma, J. Zhang, Z. Li, L. Qi, Q. Yu, and Y. Shi, "Steady progress beats stagnation: Mutual aid of foundation and conventional models in mixed domain semi-supervised medical image segmentation," in *CVPR*, 2025, pp. 5175–5185.

[21] Z. Cheng, Q. Wei, H. Zhu, Y. Wang, L. Qu, W. Shao, and Y. Zhou, "Unleashing the potential of SAM for medical adaptation via hierarchical decoding," in *CVPR*, 2024, pp. 3511–3522.

[22] X. Lin, Y. Xiang, L. Yu, and Z. Yan, "Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting," in *MICCAI*. Springer, 2024, pp. 24–34.

[23] X. Lin, Y. Xiang, Z. Wang, K.-T. Cheng, Z. Yan, and L. Yu, "SAMCT: Segment any CT allowing labor-free task-indicator prompts," *IEEE Trans. Med. Imaging*, vol. 44, no. 3, pp. 1386–1399, 2025.

[24] Z.-H. Bo, H. Qiao, C. Tian, Y. Guo, W. Li, T. Liang, D. Li, D. Liao *et al.*, "Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network," *Patterns*, vol. 2, no. 2, 2021.

[25] K. M. Timmins, I. C. van der Schaaf, E. Bennink, Y. M. Ruigrok, X. An, M. Baumgartner, P. Bourdon, R. De Feo, T. Di Noto, F. Dubost *et al.*, "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: the ADAM challenge," *NeuroImage*, vol. 238, p. 118216, 2021.

[26] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[27] G. Luo, M. Xu, H. Chen, X. Liang, X. Tao, D. Ni, H. Jeong, C. Kim, R. Stock, M. Baumgartner *et al.*, "Tumor detection, segmentation and classification challenge on automated 3D breast Ultrasound: The TDSC-ABUS challenge," *arXiv preprint arXiv:2501.15588*, 2025.

[28] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, vol. 5. Munich, Germany, 2015, p. 12.
[29] A. Degerli, S. Kiranyaz, M. E. Chowdhury, and M. Gabbouj, "Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images," in *ICIP*. IEEE, 2022, pp. 2306–2310.
[30] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *CVPR*, 2022, pp. 11 976–11 986.
[31] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt v2: Co-designing and scaling convnets with masked autoencoders," in *CVPR*, 2023, pp. 16 133–16 142.
[32] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *ICCV*, 2017, pp. 764–773.
[33] Z. Li, H. Pan, Y. Zhu, and A. K. Qin, "PGD-UNet: A position-guided deformable network for simultaneous segmentation of organs and tumors," in *IJCNN*, 2020, pp. 1–8.
[34] J. Chen, X. Wang, Z. Guo, X. Zhang, and J. Sun, "Dynamic region-aware convolution," in *CVPR*, 2021, pp. 8064–8073.
[35] H. Wang, K.-N. Wang, J. Hua, Y. Tang, Y. Chen, G.-Q. Zhou, and S. Li, "Dynamic spectrum-driven hierarchical learning network for polyp segmentation," *Med. Image Anal.*, vol. 101, p. 103449, 2025.
[36] D. Dai, C. Dong, H. Huang, F. Liu, Z. Li, and S. Xu, "Improving the performance of medical image segmentation with instructive feature learning," *Med. Image Anal.*, vol. 107, p. 103818, 2026.
[37] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *CVPR*, 2023, pp. 6185–6194.
[38] Z. Ding, Y. Zhang, C. Zhu, G. Zhang, X. Li, N. Jiang, Y. Que, Y. Peng, and X. Guan, "CAT-Unet: An enhanced u-net architecture with coordinate attention and skip-neighborhood attention transformer for medical image segmentation," *Inf. Sci.*, vol. 670, p. 120578, 2024.
[39] D.-X. Huang, X.-H. Zhou, X.-L. Xie, S.-Q. Liu, Z.-Q. Feng, M.-J. Gui, H. Li, T.-Y. Xiang, X.-L. Liu, and Z.-G. Hou, "MOSformer: Momentum encoder-based inter-slice fusion transformer for medical image segmentation," *CoRR*, vol. abs/2401.11856, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.11856
[40] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3D UX-Net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation," in *ICLR*, 2023.
[41] M. M. Rahman, S. Shokouhmand, S. Bhatt, and M. Faezipour, "MIST: Medical image segmentation transformer with convolutional attention mixing (cam) decoder," in *WACV*, 2024, pp. 404–413.
[42] J. Yang, P. Qiu, Y. Zhang, D. S. Marcus, and A. Sotiras, "D-net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation," *Biomed. Signal Process. Control*, vol. 113, p. 108837, 2026.
[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "LoRA: Low-rank adaptation of large language models." in *ICLR*, 2022.
[44] S. Jie and Z.-H. Deng, "FacT: Factor-tuning for lightweight adaptation on vision transformer," in *AAAI*, vol. 37, no. 1, 2023, pp. 1060–1068.
[45] Z. Wei, W. Dong, P. Zhou, Y. Gu, Z. Zhao, and Y. Xu, "Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation," in *MICCAI*. Springer, 2024, pp. 533–543.
[46] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *CVPR*, 2019, pp. 9308–9316.
[47] R. F. Woolson, "Wilcoxon signed-rank test," *Wiley Encycl. Clin. Trials*, pp. 1–3, 2007.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[49] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary iou: Improving object-centric image segmentation evaluation," in *CVPR*, 2021, pp. 15 334–15 342.



Kwanseok Oh (Member, IEEE) received the Ph.D. degree in the Department of Artificial Intelligence at Korea University, Seoul, Republic of Korea, in 2026. He is currently an AI Research Scientist with the Neuro AI Research Team, DEEPNOID Inc., Seoul, where he leads the development of a brain foundation model. His current research interests include medical/computer vision, machine/deep learning, domain generalization, foundation model, explainable AI, and medical image segmentation.

Dr. Oh was awarded the NAVER Ph.D. Fellowship Award in 2022, was a Finalist for the Qualcomm Innovation Fellowship Korea in 2022, and received the KU Achievement Award as the best student in his department from Korea University in 2023. He was also selected as a recipient of the 4th JW Foundation Fundamental Scientist Scholarship in 2023. His team ranked 1st in CTA segmentation, and 2nd in MRA segmentation in the TopBrain 2025 Challenge and 2nd in detection and 3rd in segmentation in the TopCoW 2024 Challenge, both held at MICCAI. He has served as a reviewer for journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI), IEEE Transactions on Medical Imaging (IEEE TMI), IEEE Transactions on Neural Networks and Learning Systems (IEEE TNNLS), Medical Image Analysis, MICCAI, *etc.*



Dahye Lee received the MS degree, and she is currently an AI Research Scientist with the Neuro AI Research Team, DEEPNOID Inc., Seoul. Her current research interests include medical/computer vision, deep learning, foundation model, and medical image segmentation.

Her team ranked 1st in CTA segmentation, and 2nd in MRA segmentation in the TopBrain 2025 Challenge and 2nd in detection and 3rd in segmentation in the TopCoW 2024 Challenge, both held at MICCAI.



Heung-Il Suk (Senior Member, IEEE) was a Visiting Professor with the Department of Radiology, Duke University, Durham, NC, USA, from 2022 to 2023. He is currently a Professor with the Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea, where he has also served as the Deputy Vice President for Research Affairs since 2025. His research interests include machine/deep learning, explainable artificial intelligence (AI), biomedical data analysis, and brain–computer interface.

Dr. Suk was awarded a Kakao Faculty Fellowship in 2018, a Young Researcher Award from the Korean Society for Human Brain Mapping (KHBM) in 2019, and an Award of Contribution to the Development of the Innovative Medical Device Industry from the Korean Ministry of Food and Drug Safety in 2023. He currently serves as an Associate Editor for the IEEE Transactions on Medical Imaging (IEEE TMI), Clinical and Molecular Hepatology (Artificial Intelligence Editor), Frontiers in Radiology (Artificial Intelligence in Radiology), Frontiers in Human Neuroscience (Brain–Computer Interface), Electronics, the International Journal of Imaging Systems and Technology (IJIST), and Biomedical Research and Clinical Practice (BRCP), and as an Area Editor for the KSII Transactions on Internet and Information Systems. He has also served as a Program Committee Member or a reviewer for NeurIPS, ICML, ICLR, AAAI, IJCAI, CVPR, ECCV, MICCAI, *etc.*