SAGEPROTO: MITIGATING CLASS-WISE LEARNING DISPARITY IN SEMI-SUPERVISED MULTI-ORGAN SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image segmentation requires dense expert annotations, motivating semi-supervised learning with limited labeled and abundant unlabeled data. However, anatomical classes differ markedly in learning difficulty, creating pronounced class-wise learning disparity under limited supervision: well-learned classes form reliable visual representations, whereas difficult classes remain under-optimized. In semi-supervised learning, unlabeled supervision and class representation learning remain closely tied to model predictions and visual evidence. These class-wise disparities may therefore persist throughout training. We propose **SageProto**, a semantic-anchor prototype framework for semi-supervised multi-organ segmentation that introduces textual semantic references into representation learning, prototype estimation, and decoding. SageProto encodes anatomical descriptions into textual semantic anchors that provide class-specific references independent of visual representation strength. These anchors modulate visual representations and initialize class prototypes, which are iteratively rectified with visual evidence and used as class-specific queries for textual and visual retrieval during decoding. Prototype regularization further constrains cross-view consistency, class semantics, and inter-class geometry. By reducing difficult-class dependence on weak visual evidence, SageProto mitigates class-wise learning disparity. Extensive experiments on two multi-organ segmentation benchmarks involving 11 semi-supervised methods show that SageProto improves the best existing Dice by 2.08% on Synapse and 2.31% on FLARE22. Code is available at SageProto.

1 INTRODUCTION

Multi-organ medical image segmentation delineates anatomical structures for quantitative analysis, treatment planning, and surgical navigation (Huang et al., 2025), yet obtaining dense annotations requires substantial clinical expertise and labeling effort (Luo et al., 2021b). Semi-supervised learning (SSL) reduces this burden by exploiting limited labeled and abundant unlabeled data (Chen et al., 2023). However, anatomical structures differ in appearance, spatial extent, and discriminative cues, leading to substantial variation in class-wise learning difficulty (Xu et al., 2025a; Zhao et al., 2024). Under limited supervision, this variation develops into a pronounced learning disparity: well-learned classes form reliable visual representations, whereas difficult classes remain under-optimized.

Recent semi-supervised segmentation methods alleviate this disparity from different perspectives, including calibrating class-wise label distributions (Lin et al., 2022), rebalancing class-wise optimization through gradient statistics (Qi et al., 2024), and reducing the distribution mismatch between labeled and unlabeled data through generative dual-distribution modeling (Huang et al., 2026). However, their learning signals remain dependent on visual evidence whose reliability varies across classes, allowing the disparity to propagate into pseudo-label supervision. Well-learned classes therefore receive increasingly reliable supervision, whereas poorly learned classes continue to receive less reliable signals, hindering learning and widening the class-wise disparity (Gu et al., 2026).

Mitigating class-wise learning disparity requires class-specific semantic information that remains reliable when visual representations are weak. Textual anatomical descriptions provide such information by characterizing each class through identity, morphology, spatial location, and relationships

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with neighboring structures (Gu et al., 2026). These attributes describe stable anatomical properties of each organ and are not determined by visual representation quality, making them valuable for classes that remain poorly represented under limited supervision. Vision-language models have demonstrated the value of textual semantics for visual representation learning (Li et al., 2026), while medical image segmentation methods incorporate anatomical knowledge through clinical context (Huemann et al., 2024), multi-scale semantic descriptions (Shan et al., 2025), and inter-organ structural information (Gu et al., 2026). These textual descriptions provide class-specific semantic cues that serve as complementary class-wise evidence when visual discriminability is limited.

Recent vision-language segmentation methods introduce semantics through pixel-text or mask-text interaction (Huang et al., 2026; Das et al., 2024), but these interactions remain conditioned on image-derived representations and may retain class-wise differences in representation quality. Prototype learning instead represents each class with an explicit class-level embedding by consolidating its associated visual features (Xu et al., 2025b; Lv et al., 2026). This formulation separates class-wise representations from individual pixel responses and provides a direct interface for incorporating class-specific semantics and modeling inter-class relations. However, existing prototypes are still estimated primarily from visual evidence, making them less reliable for difficult classes whose representations remain weak under limited supervision. This limitation motivates introducing stable semantic references into prototype construction to reduce its dependence on weak visual evidence.

We propose **SageProto**, a semantic-anchor prototype framework for semi-supervised multi-organ segmentation. SageProto introduces textual semantic anchors that provide class-specific references whose semantics are not determined by visual representation strength throughout representation learning, prototype estimation, and decoding, as illustrated in Fig. 1. Within a student-teacher architecture, Semantic-Anchor Visual Modulation introduces anatomical semantics into the input and encoder representations. Semantic-Anchor Prototype Estimation initializes class prototypes from the anchors and iteratively rectifies them with visual evidence. The rectified prototypes then act as class-specific queries in Query-Adaptive Prototype Interaction to retrieve textual and visual features across decoder layers, while Prototype Regularization constrains cross-view consistency, class semantics, and inter-class geometry. By introducing textual semantic references across representation learning, prototype estimation, and decoding, SageProto reduces the dependence of difficult classes on weak visual evidence and mitigates class-wise learning disparity. Extensive experiments on two multi-organ segmentation benchmarks involving 11 semi-supervised methods show that SageProto improves the best existing DSC by 2.08% on Synapse and 2.31% on FLARE22.

The contributions of this work are summarized as follows:

- We identify class-wise learning disparity in semi-supervised multi-organ segmentation and propose **SageProto**, a semantic-anchor prototype framework that introduces textual semantic references to reduce difficult-class dependence on weak visual evidence.
- We develop Semantic-Anchor Visual Modulation and Semantic-Anchor Prototype Estimation to incorporate textual anatomical semantics into visual representations and derive class prototypes through semantic initialization and iterative visual rectification.
- We introduce Query-Adaptive Prototype Interaction and Prototype Regularization to use rectified prototypes as class-specific queries for textual-visual retrieval across decoder layers while preserving cross-view consistency, class semantics, and inter-class geometry.

2 Related Work

Semi-Supervised Medical Image Segmentation. Semi-supervised learning (SSL) reduces annotation burden by exploiting limited labeled images and unlabeled data (Tarvainen & Valpola, 2017), with entropy minimization and consistency regularization as established paradigms (Huang et al., 2025). Entropy minimization encourages confident predictions on unlabeled samples, with later methods incorporating semantic constraints (Miao et al., 2023) and distribution regularization (Chen et al., 2026), while consistency regularization promotes prediction invariance under perturbations (Huang et al., 2025). Following these paradigms, MagicNet (Chen et al., 2023) uses volumetric partition and recovery to exchange complementary information across anatomical regions, whereas GuidedNet (Zhao et al., 2024) leverages labeled samples to guide prediction and pseudo-label learning on unlabeled data. Multi-organ segmentation further exhibits substantial variation in class-wise

learning difficulty due to differences in appearance, spatial extent, and discriminative cues (Xu et al., 2025a; Zhao et al., 2024). Recent methods alleviate this issue through class-wise label calibration (Lin et al., 2022), gradient-based optimization rebalancing (Qi et al., 2024), and generative dual-distribution modeling (Huang et al., 2026). However, their supervision remains predominantly derived from visual evidence, whose class-dependent reliability can perpetuate class-wise learning disparity during semi-supervised optimization (Gu et al., 2026).

Vision-Language Learning for Medical Image Segmentation. Vision-language learning introduces textual semantics into visual representations through image-text learning (Radford et al., 2021). MedCLIP (Wang et al., 2022) extends this paradigm to medical image-text learning. For segmentation, LViT (Li et al., 2023) incorporates text annotations into visual features, ConTEXTual-Net (Huemann et al., 2024) integrates radiology-report embeddings, and STPNet (Shan et al., 2025) retrieves semantic descriptions for lesion segmentation. Pixel-text and mask-text interactions further incorporate textual semantics: CLIP-S4 (He et al., 2023) associates pixel-wise features with textual embeddings, MTA-CLIP (Das et al., 2024) extends semantic interaction to mask representations, and TAK (Gu et al., 2026) incorporates organ morphology and inter-organ positional descriptions. However, these interactions remain conditioned on image-derived representations and may preserve class-wise disparities in representations, particularly for difficult classes under limited supervision.

Prototype Learning for Medical Image Segmentation. Prototype learning encodes each class into an explicit class-level representation by consolidating class-associated visual features (Xu et al., 2025b; Lv et al., 2026). This formulation reduces sensitivity to individual pixel responses and provides a direct interface for incorporating class-specific semantics and modeling inter-class relations. HMPD (Xu et al., 2025b) constructs hierarchical prototypes to capture intra-class variation, while PCR (Xu et al., 2022) enforces prototype consistency across perturbed views. Probabilistic representations and distribution prototypes (Yuan et al., 2025) characterize uncertain pixels and ambiguous boundaries, whereas CPL (Lv et al., 2026) learns class-aware prototypes through cluster queries to improve discrimination and mitigate class imbalance. However, existing prototypes are still derived primarily from visual evidence, making prototype estimation vulnerable to weak representations of difficult classes under limited supervision. This limitation motivates class-specific textual semantic references for more reliable prototype estimation.

3 METHOD

3.1 OVERALL FRAMEWORK

Given a labeled set $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ and an unlabeled set $\mathcal{D}_u = \{x_j\}_{j=1}^{N_u}$, with $N_u \gg N_l$, we consider semi-supervised medical image segmentation through prototype learning under limited supervision. Conventional prototype estimation relies on visual evidence. When classes differ in learning difficulty, well-learned classes form prototypes from stronger visual representations, whereas poorly learned classes rely on weaker evidence. SageProto mitigates this class learning bias by introducing class-specific semantic anchors as references throughout representation learning, prototype estimation, and decoding, as illustrated in Fig. 1. SageProto derives $\widetilde{P}_c^{(0)}$ from semantic-anchor weights and rectifies it with visual evidence. It adopts a student-teacher architecture with semantic anchors and EMA-updated teacher parameters. Semantic-Anchor Visual Modulation introduces anchor semantics into image and encoder representations to form semantic-aware features. Semantic-Anchor Prototype Estimation initializes and iteratively rectifies prototypes from these features and anchors. Query-Adaptive Prototype Interaction uses rectified prototypes as class-specific queries to retrieve and integrate textual and visual features across decoder layers, while Prototype Regularization constrains cross-view consistency, class semantics, and inter-class geometry.

3.2 SEMANTIC-ANCHOR VISUAL MODULATION

Semantic-Anchor Visual Modulation derives anchors from class-specific textual descriptions and incorporates them at image and encoder levels to stabilize class representations across different learning difficulties through class-specific image modulation and hierarchical semantic injection.

Anatomical Semantic Anchors. Semantic anchors provide class-specific textual semantics independent of visual statistics. For each class $c \in \{1, \ldots, C\}$, the MedCPT query encoder $\Phi_{\text{text}}(\cdot)$

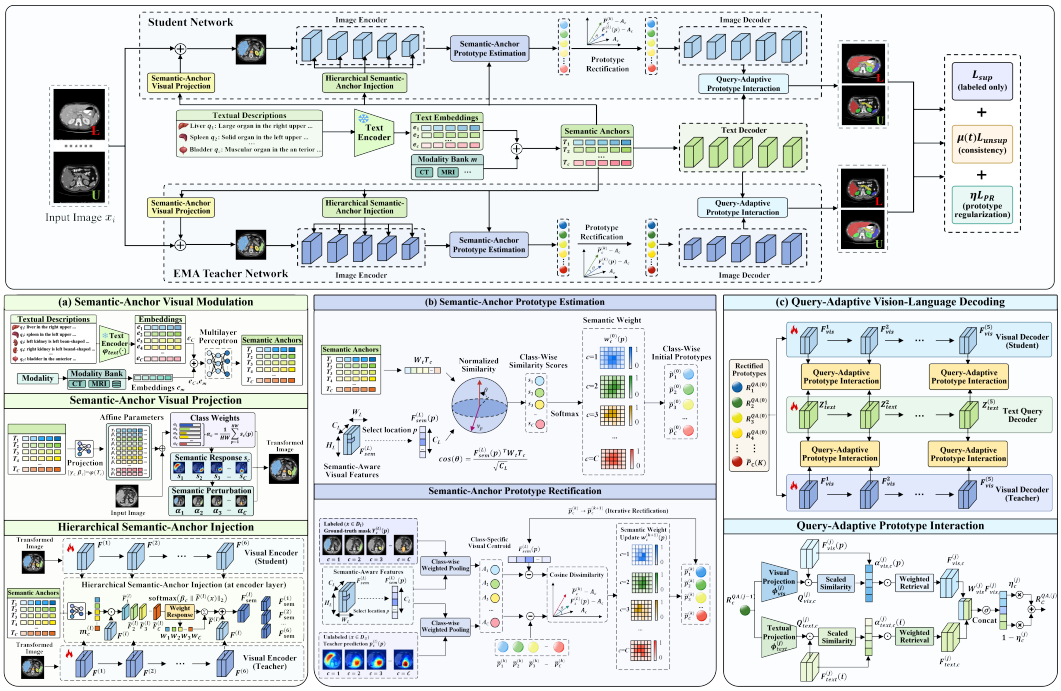


Figure 1: Overall framework of **SageProto**: Semantic-Anchor Visual Modulation, Semantic-Anchor Prototype Estimation, Query-Adaptive Vision-Language Decoding, and Prototype Regularization.

maps the anatomical description q_c to $e_c \in \mathbb{R}^{768}$, while the modality embedding matrix E_{mod} maps imaging modality m to $e_m \in \mathbb{R}^{768}$. The semantic anchor $T_c \in \mathbb{R}^d$ is obtained by projecting $e_c + e_m$ through $g_\phi(\cdot)$, implemented with two MLP layers. The anchors form $T = [T_1, \ldots, T_C]^\top \in \mathbb{R}^{C \times d}$:

$$e_c = \Phi_{\text{text}}(q_c), \quad e_m = E_{\text{mod}}(m), \quad T_c = g_\phi\left(e_c + e_m\right). \tag{1}$$

Semantic-Anchor Visual Projection. Semantic-Anchor Visual Projection introduces class-specific semantics into the input while adapting their influence to the current image. Given $I \in \mathbb{R}^{C_{\text{in}} \times H \times W}$, a two-layer projection $\Psi(\cdot)$ maps each semantic anchor $T_c \in \mathbb{R}^d$ to channel-wise scale and shift parameters $[\gamma_c, \beta_c] = \Psi(T_c) \in \mathbb{R}^{2C_{\text{in}}}$. Image-specific class relevance is captured by an image-level weight α_c, estimated from the correspondence between the input and the projected anchor $W_a T_c$, where $W_a \in \mathbb{R}^{C_{\text{in}} \times d}$. Class-specific affine residuals are weighted by α_c and incorporated into the input, so that relevant semantic anchors modulate its channel-wise representation:

$$\alpha_c = \frac{1}{HW} \sum_{p=1}^{HW} \text{softmax}_c\left(\frac{I(p)^\top W_a T_c}{\sqrt{C_{\text{in}}}}\right), \quad I_{\text{sct}} = I + \alpha_{\text{sct}} \sum_{c=1}^{C} \alpha_c \left[(\gamma_c - 1) \odot I + \beta_c\right]. \tag{2}$$

where p indexes spatial locations, $\text{softmax}_c(\cdot)$ normalizes responses across classes, and $\odot$ denotes channel-wise multiplication with spatial broadcasting. The scalar α_{sct} controls modulation strength.

Hierarchical Semantic-Anchor Injection. Hierarchical Semantic-Anchor Injection introduces class-specific textual semantics throughout the visual encoder. At encoder layer l, a two-layer projection $\Psi_m^{(l)}(\cdot)$ maps each semantic anchor T_c to a channel-wise response that modulates $F^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$, producing $\hat{F}_c^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$. Spatially averaged feature magnitudes define class-wise weights for constructing $\bar{F}^{(l)}$, which is interpolated with $F^{(l)}$ to obtain $F_{\text{sem}}^{(l)} \in \mathbb{R}^{C_l \times H_l \times W_l}$, preserving visual information while introducing class-specific semantics across encoder stages.

$$\hat{F}_c^{(l)} = F^{(l)} \odot \left[1 + \sigma\left(\Psi_m^{(l)}(T_c)\right)\right], \tag{3}$$

$$\bar{F}^{(l)} = \sum_{c=1}^{C} \text{softmax}_c\left(\mathbb{E}_p\left[\left\|\hat{F}_c^{(l)}(p)\right\|_2\right]\right) \hat{F}_c^{(l)}, \tag{4}$$

$$F_{\text{sem}}^{(l)} = (1-\rho)F^{(l)} + \rho\bar{F}^{(l)}, \tag{5}$$

where p indexes locations, $\mathbb{E}_p[\cdot]$ denotes spatial averaging, $\sigma(\cdot)$ denotes sigmoid, $\mathrm{softmax}_c(\cdot)$ normalizes responses across classes, and $\odot$ denotes channel-wise multiplication with spatial broadcasting. The scalar $\rho \in [0,1]$ adaptively balances the semantic and original representations.

3.3 Semantic-Anchor Prototype Estimation

Semantic-Anchor Prototype Estimation integrates textual and visual evidence to construct reliable prototypes across learning difficulties through semantic initialization and iterative rectification.

Semantic-Anchor Prototype Initialization. Textual semantic anchors provide references for estimating prototypes from semantic-aware visual features. At the final encoder layer L, $F_{\text{sem}}^{(L)}(p) \in \mathbb{R}^{C_L}$ denotes the semantic-aware features $F_{\text{sem}}^{(L)}$ at location p. For each class $c \in \{1,\dots,C\}$, the semantic anchor $T_c \in \mathbb{R}^d$ is projected into the visual space by $W_t \in \mathbb{R}^{C_L \times d}$. The semantic weight $\omega_c^{(0)}(p) \in [0,1]$ and initial semantic-anchor prototype $\widetilde{P}_c^{(0)} \in \mathbb{R}^{C_L}$ are defined as:

$$\omega_c^{(0)}(p) = \mathrm{softmax}_c\left(\frac{F_{\text{sem}}^{(L)}(p)^\top W_t T_c}{\sqrt{C_L}}\right), \quad \widetilde{P}_c^{(0)} = \frac{\sum_{p=1}^N \omega_c^{(0)}(p) F_{\text{sem}}^{(L)}(p)}{\sum_{p=1}^N \omega_c^{(0)}(p) + \epsilon}, \tag{6}$$

where $N = H_L W_L$ is the number of spatial locations, $\mathrm{softmax}_c(\cdot)$ normalizes similarity across classes, $\sqrt{C_L}$ provides dimensional scaling, and $\epsilon = 10^{-8}$ ensures numerical stability.

Semantic-Anchor Prototype Rectification. Semantic-Anchor Prototype Rectification uses class-specific visual centroids to iteratively rectify anchor-derived semantic weights. The class-specific visual centroid $A_c \in \mathbb{R}^{C_L}$ is estimated from reliable class evidence. For labeled images, the ground-truth mask $Y_c^{(L)}$ is used, while for unlabeled images, the EMA teacher prediction $p_c^{(t,L)}(p)$, spatially aligned with the final encoder resolution, provides class-specific evidence:

$$A_c = \begin{cases} \dfrac{\sum_{p=1}^N Y_c(p) F_{\text{sem}}^{(L)}(p)}{\sum_{p=1}^N Y_c(p) + \epsilon}, & x \in \mathcal{D}_l, \\ \dfrac{\sum_{p=1}^N p_c^{(t,L)}(p) F_{\text{sem}}^{(L)}(p)}{\sum_{p=1}^N p_c^{(t,L)}(p) + \epsilon}, & x \in \mathcal{D}_u, \end{cases} \tag{7}$$

where $N = H_L W_L$ and $\epsilon = 10^{-8}$. Rectification is applied only to classes with reliable evidence in the image, while the remaining prototypes are kept unchanged.

At each rectification step k, the semantic-aware feature $F_{\text{sem}}^{(L)}(p)$ and the semantic-anchor prototype $\widetilde{P}_c^{(k)}$ are centered by A_c. Their directional inconsistency is measured by cosine dissimilarity $d_c^{(k)}(p)$:

$$d_c^{(k)}(p) = 1 - \cos\left(F_{\text{sem}}^{(L)}(p) - A_c, \widetilde{P}_c^{(k)} - A_c\right). \tag{8}$$

The semantic weight $\omega_c^{(k+1)}(p)$ is updated from $\omega_c^{(k)}(p)$ using rectification strength $\lambda_c \geq 0$, and the semantic-anchor prototype $\widetilde{P}_c^{(k+1)}$ is re-estimated from the updated weights:

$$\omega_c^{(k+1)}(p) = \omega_c^{(k)}(p) \exp\left(-\lambda_c d_c^{(k)}(p)\right), \quad \widetilde{P}_c^{(k+1)} = \frac{\sum_{p=1}^N \omega_c^{(k+1)}(p) F_{\text{sem}}^{(L)}(p)}{\sum_{p=1}^N \omega_c^{(k+1)}(p) + \epsilon}. \tag{9}$$

$\widetilde{P}_c^{(k+1)} = \widetilde{P}_c^{(k)}$. After K steps, $\widetilde{P}_c^{(K)} \in \mathbb{R}^{C_L}$ defines the semantic-anchor prototype.

3.4 Query-Adaptive Prototype Interaction

Query-Adaptive Prototype Interaction (QAPI) initializes class-specific queries from the rectified semantic-anchor prototype $\widetilde{P}_c^{(K)} \in \mathbb{R}^{C_L}$ and progressively updates them across decoder layers. We define $R_c^{\text{QA},(0)} = \widetilde{P}_c^{(K)}$. At decoder layer $j \in \{1,\dots,J\}$, learnable projections $\Phi_{\text{text}}^{(j)}(\cdot)$ and

$\Phi_{\text{vis}}^{(j)}(\cdot)$ map $R_c^{\text{QA},(j-1)}$ from its layer-specific input dimension, C_L for $j=1$ and D_{j-1} for $j>1$, into textual and visual queries $Q_{\text{text},c}^{(j)} \in \mathbb{R}^{D_{\text{text}}^{(j)}}$ and $Q_{\text{vis},c}^{(j)} \in \mathbb{R}^{D_{\text{vis}}^{(j)}}$:

$$Q_{\text{text},c}^{(j)} = \Phi_{\text{text}}^{(j)}\left(R_c^{\text{QA},(j-1)}\right), \quad Q_{\text{vis},c}^{(j)} = \Phi_{\text{vis}}^{(j)}\left(R_c^{\text{QA},(j-1)}\right). \tag{10}$$

The textual and visual decoders produce $F_{\text{text}}^{(j)}(i)$ and $F_{\text{vis}}^{(j)}(p)$ from semantic-anchor-injected textual features and the corresponding visual representation, where i and p index textual embeddings and spatial locations, respectively. The prototype-conditioned queries then derive class-specific weights $\alpha_{\text{text},c}^{(j)}(i)$ and $\alpha_{\text{vis},c}^{(j)}(p)$ through scaled dot-product similarity:

$$\alpha_{\text{text},c}^{(j)}(i) = \text{softmax}_i\left(\frac{F_{\text{text}}^{(j)}(i)^\top Q_{\text{text},c}^{(j)}}{\sqrt{D_{\text{text}}^{(j)}}}\right), \quad \alpha_{\text{vis},c}^{(j)}(p) = \text{softmax}_p\left(\frac{F_{\text{vis}}^{(j)}(p)^\top Q_{\text{vis},c}^{(j)}}{\sqrt{D_{\text{vis}}^{(j)}}}\right). \tag{11}$$

The prototype-derived weights retrieve class-specific representations from the textual and visual features, $F_{\text{text},c}^{(j)} \in \mathbb{R}^{D_{\text{text}}^{(j)}}$ and $F_{\text{vis},c}^{(j)} \in \mathbb{R}^{D_{\text{vis}}^{(j)}}$, for cross-modal interaction:

$$F_{\text{text},c}^{(j)} = \sum_{i=1}^{N_{\text{text}}^{(j)}} \alpha_{\text{text},c}^{(j)}(i) F_{\text{text}}^{(j)}(i), \quad F_{\text{vis},c}^{(j)} = \sum_{p=1}^{N_{\text{vis}}^{(j)}} \alpha_{\text{vis},c}^{(j)}(p) F_{\text{vis}}^{(j)}(p). \tag{12}$$

QAPI projects the textual and visual representations into a shared D_j-dimensional space, where their concatenated projections are passed through a linear layer followed by a sigmoid function to obtain $\eta_c^{(j)} \in [0,1]$. The textual and visual representations are then adaptively fused as

$$R_c^{\text{QA},(j)} = \eta_c^{(j)} W_{\text{vis}}^{(j)} F_{\text{vis},c}^{(j)} + \left(1-\eta_c^{(j)}\right) W_{\text{text}}^{(j)} F_{\text{text},c}^{(j)}, \tag{13}$$

where $W_{\text{text}}^{(j)}$ and $W_{\text{vis}}^{(j)}$ are learnable projection matrices, and $\eta_c^{(j)} \in [0,1]$ is predicted from the concatenated projected features. For $j < J$, $R_c^{\text{QA},(j)}$ is propagated to the next decoder layer to generate new textual and visual queries through Eq. 10, enabling progressive class-specific interaction.

3.5 Prototype Regularization

Prototype Regularization stabilizes semantic-anchor prototypes across learning difficulties through cross-view consistency $\mathcal{L}\text{cons}$, semantic consistency $\mathcal{L}\text{text}$, and inter-class geometry $\mathcal{L}_{\text{geo}}$, with λ_1 and λ_2 balancing semantic and geometric regularization strengths:

$$\mathcal{L}_{\text{PR}} = \mathcal{L}_{\text{cons}} + \lambda_1 \mathcal{L}_{\text{text}} + \lambda_2 \mathcal{L}_{\text{geo}}, \tag{14}$$

Cross-View Prototype Consistency. Cross-View Prototype Consistency reduces student-teacher prototype variation with sparse evidence. It minimizes cosine dissimilarity between $\widetilde{P}_{c,\text{s}}^{(K)}$ and $\widetilde{P}_{c,\text{t}}^{(K)}$ for $c \in \mathcal{C}_v$, where $\mathcal{C}_v$ includes classes valid across branches and $\text{sg}(\cdot)$ blocks teacher gradients:

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\mathcal{C}_v|} \sum_{c \in \mathcal{C}_v} \left[1 - \cos\left(\widetilde{P}_{c,\text{s}}^{(K)}, \text{sg}\left(\widetilde{P}_{c,\text{t}}^{(K)}\right)\right)\right]. \tag{15}$$

Semantic Prototype Consistency. Semantic Prototype Consistency preserves semantics by aligning the student prototype $\widetilde{P}_{c,\text{s}}^{(K)} \in \mathbb{R}^{C_L}$ with its projected anchor $W_t T_c \in \mathbb{R}^{C_L}$ through cosine similarity, encouraging the rectified prototype to maintain alignment with the anchor semantics:

$$\mathcal{L}_{\text{text}} = \frac{1}{|\mathcal{C}_v|} \sum_{c \in \mathcal{C}_v} \left[1 - \cos\left(\widetilde{P}_{c,\text{s}}^{(K)}, W_t T_c\right)\right]. \tag{16}$$

Prototype Geometry Regularization. Prototype Geometry Regularization preserves inter-class relations encoded by semantic anchors. For class pairs $\mathcal{P}_v = \{(c,k) \mid c,k \in \mathcal{C}_v, c<k\}$, prototype and semantic relations are measured by cosine similarity, with semantic relations detached:

$$G_{c,k}^{\text{proto}} = \cos\big(\widetilde{P}_{c,\text{s}}^{(K)}, \widetilde{P}_{k,\text{s}}^{(K)}\big), \quad G_{c,k}^{\text{sem}} = \cos\big(\text{sg}(W_t T_c), \text{sg}(W_t T_k)\big). \tag{17}$$

The geometric objective $\mathcal{L}_{\text{geo}}$ aligns prototype relations with the corresponding semantic relations:

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\mathcal{P}_v|} \sum_{(c,k) \in \mathcal{P}_v} \left(G_{c,k}^{\text{proto}} - G_{c,k}^{\text{sem}}\right)^2. \tag{18}$$

3.6 Optimization Objective

SageProto adopts a student-teacher framework for multi-class semi-supervised segmentation, with the teacher updated by exponential moving average (EMA). For labeled input x_i, the student predicts $p_i^{(s)}$. The supervised objective combines cross-entropy, Dice, and boundary losses, where y_i is the ground-truth mask, α balances cross-entropy and Dice, and λ_b weights the boundary term:

$$\mathcal{L}_{\text{sup}} = (1-\lambda_b)\left[\alpha\mathcal{L}_{\text{CE}}\left(p_i^{(s)}, y_i\right) + (1-\alpha)\mathcal{L}_{\text{DSC}}\left(p_i^{(s)}, y_i\right)\right] + \lambda_b\mathcal{L}_{\text{bd}}\left(p_i^{(s)}, y_i\right), \tag{19}$$

For unlabeled input x_j, student-teacher consistency is enforced at confident locations. At location u, $p_j^{(s)}(u)$ and $p_j^{(t)}(u)$ are probability vectors over C classes, with teacher confidence $s_j(u) = \max_c p_{j,c}^{(t)}(u)$. Given threshold $\tau(t)$, $\mathbb{I}(\cdot)$ selects locations satisfying $s_j(u) \geq \tau(t)$. The unsupervised objective minimizes the squared ℓ_2 distance between student and teacher predictions:

$$\mathcal{L}_{\text{unsup}}^{(j)} = \frac{\sum_u \mathbb{I}(s_j(u) \geq \tau(t)) \left\| p_j^{(s)}(u) - p_j^{(t)}(u) \right\|_2^2}{C\left(\sum_u \mathbb{I}(s_j(u) \geq \tau(t)) + \epsilon\right)}, \quad \epsilon = 10^{-8}. \tag{20}$$

The overall objective combines supervised segmentation, prediction consistency, and prototype regularization, where $\mu(t)$ follows a temporal schedule and η weights prototype regularization:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \mu(t)\mathcal{L}_{\text{unsup}} + \eta\mathcal{L}_{\text{PR}}. \tag{21}$$

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics. We evaluate SageProto on two multi-organ segmentation datasets: Synapse, and FLARE22. (1) Synapse (Landman et al., 2015) contains 30 abdominal CT cases with 3,779 annotated slices covering eight organs: aorta, gallbladder (Gal), right kidney (R.Kid), left kidney (L.Kid), liver, pancreas (Pan), spleen, and stomach (Sto). Following (Li et al., 2025), 18 cases are used for training and 12 for testing. (2) FLARE22 (Ma et al., 2024) contains 13 organ classes: liver, spleen, Pan, R.Kid, L.Kid, Sto, Gal, esophagus (Eso), aorta, inferior vena cava (IVC), right adrenal gland (RAG), left adrenal gland (LAG), and duodenum (Duo). The data are partitioned into training, validation, and testing sets at a 6:2:2 ratio following (Zhao et al., 2024). We report DSC, Jaccard (Jac), Average Surface Distance (ASD), and parameter count (Params).

Implementation Details. Axial medical slices are normalized to $[0, 1]$. FLARE22 is resized to 512×512, while Synapse retains its native resolution with zero-padding to multiples of 32. Synapse and FLARE22 contain 8 and 13 foreground organs and are trained for 30k and 80k iterations, respectively. SageProto follows a Mean Teacher framework with a U-Mamba student and EMA teacher. We use SGD with momentum 0.9, weight decay 10^{-4}, and polynomial decay of power 0.9 after 500 warm-up iterations, with initial learning rates of 1×10^{-2} and 5×10^{-4}. The CE–DSC weight α is set to 0.50 and 0.36, while λ_b is 0.02 and 0.03, activated at 3k and 5k iterations. Gradients are unclipped for Synapse and clipped to norm 12 for FLARE22. EMA decay is 0.99 and $0.997 \rightarrow 0.9995$, respectively. Module schedules and additional hyperparameters are provided in Appendix 6.2.

4.2 Experimental Analysis

Comparison with State-of-the-Art Methods. As shown in Table 1, SageProto achieves the best performance on Synapse and FLARE22 under 10% labeled supervision, reaching 86.63% DSC on FLARE22, 2.74 points above SemiGDA. Larger gains are observed for difficult classes, including the gallbladder and pancreas on Synapse and the duodenum on FLARE22, while well-learned classes improve less. This class-wise pattern matches our motivation: well-learned classes receive stronger visual evidence during semi-supervised optimization, whereas difficult classes receive less reliable supervision. SageProto introduces class-specific semantic references throughout representation learning, prototype construction, and decoding, reducing the dependence of difficult classes on weak visual evidence and improving their segmentation.

Ablation Study. Table 2 analyzes the contribution of each SageProto stage. Introducing semantic anchors into visual representation learning improves DSC from 54.82% to 59.47%, while applying

Table 1: Comparison with state-of-the-art semi-supervised methods on Synapse, FLARE22, and AMOS under 10% labeled supervision. Δ denotes the relative percentage change with respect to SemiGDA. Organ-wise results are Dice (%) across all three benchmarks.

Method	Venue	with 10% Labeled Data on the Synapse Dataset								Average	
		R.Kid.	L.Kid.	Liver	Spleen	Sto.	Aorta	Gal.	Pan.	DSC (%)	ASD (mm)
MT (Tarvainen & Valpola, 2017)	NeurIPS	25.95	39.28	73.49	39.76	25.65	52.57	21.09	9.44	35.90	34.13
FixMatch (Sohn et al., 2020)	NeurIPS	45.77	53.20	86.34	67.33	41.87	75.16	21.98	18.91	51.32	20.57
CPS (Chen et al., 2021)	CVPR	26.73	40.53	77.17	39.47	22.09	51.62	20.72	9.37	35.96	33.02
BCP (Bai et al., 2023)	CVPR	65.25	62.29	76.49	68.90	43.84	68.25	28.72	23.76	54.69	38.94
CML (Wu et al., 2024)	ACMMM	51.48	58.32	79.30	56.71	12.04	74.39	9.84	9.84	43.99	44.84
GA (Qi et al., 2024)	ECCV	68.45	68.63	83.80	52.96	52.46	71.62	**40.27**	30.59	58.60	8.59
DCMamba (Li et al., 2025)	TMI	79.78	76.54	**89.80**	75.82	52.97	69.52	15.83	23.03	60.41	8.96
DiffMatch (Sun et al., 2025)	ICLR	72.58	69.64	83.68	70.77	55.59	77.16	32.10	30.35	61.49	8.10
ConformalSAM (Chen et al., 2025)	ICCV	77.61	69.11	85.10	75.77	**61.00**	74.40	33.20	30.05	63.28	6.47
DualFete (Yi et al., 2026)	AAAI	75.17	70.08	82.93	72.97	55.97	78.89	33.63	30.57	62.53	7.43
SemiGDA (Huang et al., 2026)	CVPR	73.41	73.81	82.51	74.80	58.00	81.37	34.78	31.46	63.77	7.03
SageProto (Ours)	–	**80.78**	**77.67**	76.95	**76.37**	59.72	**83.03**	38.36	**33.91**	**65.85**	**6.44**
Δ vs. SemiGDA		**↑10.04%**	**↑5.23%**	**↓6.74%**	**↑2.10%**	**↑2.97%**	**↑2.04%**	**↑10.29%**	**↑7.79%**	**↑3.26%**	**↓8.39%**

Method	Venue	with 10% Labeled Data on the FLARE22 Dataset													Average	
		Spleen	R.Kid.	L.Kid.	Liver	Sto.	Gal.	Eso.	Aorta	IVC	Pan.	RAG	LAG	Duo.	DSC (%)	Jac. (%)
MT (Tarvainen & Valpola, 2017)	NeurIPS	91.54	92.77	93.78	96.49	74.64	73.49	70.66	92.17	82.71	69.23	61.88	41.26	66.68	77.49	66.45
DTC (Luo et al., 2021a)	AAAI	92.91	92.40	92.68	96.63	72.76	78.38	69.39	91.87	81.47	66.82	59.74	59.10	65.76	78.45	67.05
CPS (Chen et al., 2021)	CVPR	92.16	92.71	92.70	96.62	77.02	75.12	72.78	92.25	81.91	69.39	63.56	58.96	65.94	79.32	68.14
MagicNet (Chen et al., 2023)	CVPR	88.04	92.95	92.18	97.04	81.51	84.36	77.07	91.75	81.01	71.15	63.34	60.33	69.61	80.79	70.23
GA (Qi et al., 2024)	ECCV	96.79	94.56	91.63	96.41	84.17	74.01	81.72	91.30	82.78	58.32	82.91	**83.09**	52.02	82.28	72.17
GuidedNet (Zhao et al., 2024)	ACMMM	93.48	93.48	**94.51**	96.77	83.19	**85.87**	75.74	92.95	84.63	75.97	71.10	67.77	**71.92**	83.64	73.08
DiffMatch (Sun et al., 2025)	ICLR	96.31	94.18	91.31	96.25	81.44	74.80	80.79	91.41	82.89	48.12	**83.68**	82.06	42.85	80.47	70.32
ConformalSAM (Chen et al., 2025)	ICCV	96.47	94.72	91.16	96.58	84.14	74.89	**81.81**	92.62	84.78	68.14	83.60	82.79	56.73	83.73	73.75
DualFete (Yi et al., 2026)	AAAI	96.49	93.61	91.46	96.29	84.21	71.90	79.55	91.19	83.56	61.20	82.59	81.10	53.28	82.03	71.68
SemiGDA (Huang et al., 2026)	CVPR	96.45	94.80	91.66	96.61	84.82	75.08	81.26	92.36	85.60	70.98	83.53	82.32	60.74	84.32	74.45
SageProto (Ours)	–	**98.12**	**96.43**	94.46	**97.52**	**88.92**	82.01	74.36	**94.33**	**87.63**	**76.98**	83.61	82.13	69.65	**86.63**	**77.67**
Δ vs. SemiGDA		**↑1.73%**	**↑1.72%**	**↑3.05%**	**↑0.94%**	**↑4.83%**	**↑9.23%**	**↓8.49%**	**↑2.13%**	**↑2.37%**	**↑8.45%**	**↑0.10%**	**↓0.23%**	**↑14.67%**	**↑2.74%**	**↑4.33%**

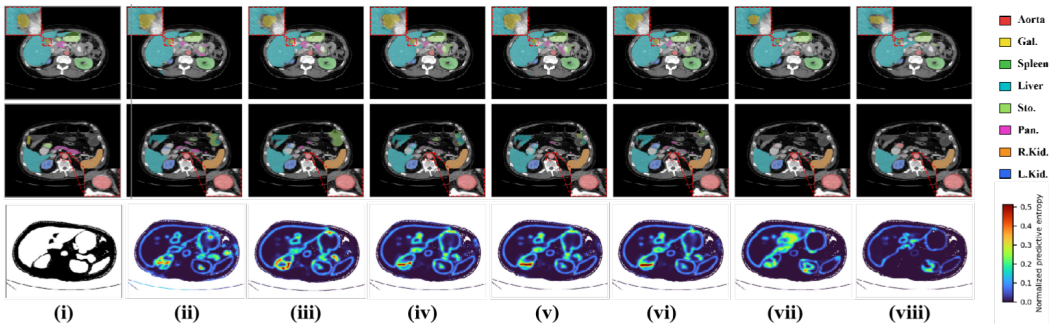


Figure 2: Qualitative comparison of segmentation and uncertainty. Columns show (i) Label, (ii) Baseline, (iii) +SAVM, (iv) +SAPE, (v) +QAPI, (vi) SAVM+SAPE, (vii) SAVM+SAPE+QAPI, and (viii) SageProto. Dashed boxes highlight challenging regions, while uncertainty contracts toward organ boundaries during training.

them to prototype estimation and query-adaptive decoding yields 60.29% and 59.91%, respectively. Using semantic anchors across all three stages raises DSC to 64.04%, indicating that class-specific semantics are most effective when maintained throughout representation learning, prototype construction, and decoding. Fig. 4 shows the class-wise learning process. Under the baseline, well-learned classes converge rapidly, whereas difficult classes remain under-optimized. SageProto produces larger gains on poorly learned classes, with the stomach improving by 29.5%, while strong classes change less. The learned feature space exhibits clearer class separation, and the negative correlation between baseline Dice and improvement ($r = -0.44$) shows that classes with lower baseline performance tend to benefit more.

Fig. 2 examines segmentation quality and predictive uncertainty. The first two rows show more complete delineation of challenging regions. Mean predictive entropy decreases from 0.069 to 0.026 on AMOS22, 0.053 to 0.027 on FLARE22, and 0.021 to 0.007 on Synapse, corresponding to reductions of 62%, 50%, and 65%. High-entropy responses gradually contract from diffuse abdominal regions toward ambiguous organ boundaries, while organ interiors become increasingly certain. These observations are consistent with the class-wise analysis: SageProto mainly improves difficult classes while reducing the gap between poorly learned and well-learned structures.

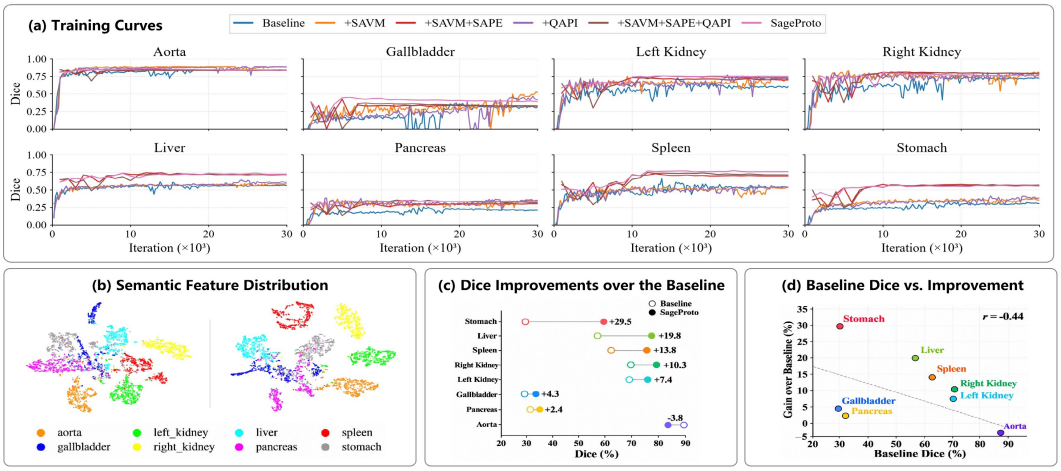


Figure 3: Training dynamics and class-wise analysis of SageProto on Synapse with 10% labeled data. (a) Dice trajectories. (b) Semantic feature distributions of baseline and SageProto. (c) Dice improvements over baseline. (d) Baseline Dice versus performance gain, showing larger gains for poorly learned classes.

Table 2: Ablation studies of SageProto on Synapse with 10% labeled data, including configurations and internal design choices. Organ-wise results are Dice (%); DSC (%) and ASSD (mm) are reported.

Model Variant	Synapse with 10% Labeled Data								Average	
	R.Kid.	L.Kid.	Liver	Spleen	Sto.	Aorta	Gal.	Pan.	DSC (%)	ASSD (mm)
(a) Overall Ablation										
Baseline	70.44	70.24	57.12	62.58	30.25	86.86	29.56	31.52	54.82	11.23
+ SAVM	80.89	70.48	58.64	53.87	35.50	88.27	52.78	35.30	59.47	9.19
+ SAPE	75.93	68.28	70.24	61.28	49.11	85.12	42.91	29.42	60.29	8.89
+ QAPI	77.40	70.58	75.95	55.84	51.80	83.14	34.78	29.76	59.91	8.79
SAVM + SAPE	73.34	71.42	76.56	75.88	61.82	80.70	34.46	29.45	62.95	7.21
SAVM + SAPE + QAPI	79.41	75.52	70.99	78.04	58.10	83.57	31.55	35.14	64.04	7.37
SageProto	80.78	77.67	76.95	76.37	59.72	83.03	33.86	33.91	**65.29**	**6.43**
Δ vs. Baseline	**↑10.34%**	**↑7.43%**	**↑19.83%**	**↑13.79%**	**↑29.47%**	**↓3.83%**	**↑4.30%**	**↑2.39%**	**↑10.47%**	**↓4.80%**
(b) Semantic Anchor Formulation										
+ Language Semantics	74.86	68.97	69.13	58.27	46.47	81.86	32.29	26.95	57.35	11.37
+ Semantic Anchors	80.89	70.48	58.64	53.87	35.50	88.27	52.78	35.30	**59.47**	**9.19**
(c) Semantic-Anchor Prototype Estimation										
Visual Prototype	79.65	69.79	70.44	59.08	54.02	85.08	33.25	29.17	60.06	9.02
Initial Prototype	77.19	72.33	70.02	72.21	56.28	83.96	32.19	29.85	61.75	7.87
Rectified Prototype	73.34	71.42	76.56	75.88	61.82	80.70	34.46	29.45	**62.95**	**7.21**
(d) Query-Adaptive Prototype Interaction										
Visual Integration	77.12	74.01	71.18	70.41	50.63	82.27	34.01	27.37	60.87	8.16
Cross-modal Integration	84.67	32.67	77.78	72.03	67.78	30.62	75.41	57.10	62.26	7.85
Query-Adaptive Integration	79.41	75.52	70.99	78.04	58.10	83.57	31.55	35.14	**64.04**	**7.37**
(e) Prototype Regularization										
Cross-View Consistency	79.41	75.52	70.99	78.04	58.10	83.57	31.55	35.14	64.04	7.65
Semantic Consistency	84.29	39.59	75.89	80.90	73.99	34.32	72.09	58.56	64.95	6.74
Prototype Geometry	80.38	76.12	74.14	70.83	57.06	84.15	38.81	34.16	64.46	6.88
Prototype Regularization	80.78	77.67	76.95	76.37	59.72	83.03	33.86	33.91	**65.29**	**6.43**

5 Conclusion

We present **SageProto**, a semantic-anchor prototype framework for semi-supervised multi-organ segmentation that mitigates class-wise learning disparity under limited supervision by introducing stable class-specific semantic references. SageProto integrates textual semantic anchors into representation learning, prototype estimation, and decoding, reducing the dependence of difficult classes on weak visual evidence. The resulting prototypes are iteratively rectified, used for textual and visual retrieval across decoder layers, and regularized by cross-view consistency, class semantics, and inter-class geometry. Experiments on two multi-organ benchmarks demonstrate consistent improvements over existing methods, with larger gains on poorly learned classes.

AI USE STATEMENT

(This section is required and does not count toward the page limit.)

In this work, we used generative AI tools for language editing, and manuscript organization. These tools were also used to assist with the interpretation and presentation of experimental results. Generative AI was not used to conduct experiments or generate reported numerical results. All AI-assisted content was reviewed by the authors. LLM-generated code was verified and tested for correctness by two authors. The methodology, experimental results, and conclusions were checked against the implemented method and original experiments. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 11514–11524, 2023.

Changrui Chen, Jungong Han, and Kurt Debattista. Exploring the stochastic regularisation in normalisation layers for semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Danhui Chen, Ziquan Liu, Chuxi Yang, Dan Wang, Yan Yan, Yi Xu, and Xiangyang Ji. Conformalsam: Unlocking the potential of foundational segmentation models in semi-supervised semantic segmentation with conformal prediction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24045–24055. IEEE, 2025.

Duowen Chen, Yunhao Bai, Wei Shen, Qingli Li, Lequan Yu, and Yan Wang. Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23869–23878, 2023.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2613–2622, 2021.

Anurag Das, Xinting Hu, Li Jiang, and Bernt Schiele. Mta-clip: Language-guided semantic segmentation with mask-text alignment. In *European Conference on Computer Vision*, pp. 39–56. Springer, 2024.

Yuliang Gu, Weilun Tsao, Yepeng Liu, Lianming Wu, Thierry Géraud, Bo Du, and Yongchao Xu. Leveraging textual anatomical knowledge for class-imbalanced semi-supervised multi-organ segmentation. *IEEE Transactions on Medical Imaging*, 2026.

Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. Clip-s4: Language-guided self-supervised semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11207–11216. IEEE, 2023.

Kaiwen Huang, Yi Zhou, Yizhe Zhang, Jingxiong Li, and Tao Zhou. Semigda: Generative dual-distribution alignment for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1450–1460, June 2026.

Wei Huang, Lei Zhang, Zizhou Wang, and Lituan Wang. Exploring inherent consistency for semi-supervised anatomical structure segmentation in medical imaging. *IEEE Transactions on Medical Imaging*, 43(11):3731–3741, 2024.

Wei Huang, Lei Zhang, Zizhou Wang, and Yan Wang. Gapmatch: Bridging instance and model perturbations for enhanced semi-supervised medical image segmentation. In *Proc. AAAI Conf. Artif. Intell.*, volume 39, pp. 17458–17466, 2025.

Zachary Huemann, Xin Tie, Junjie Hu, and Tyler J Bradshaw. Contextual net: a multimodal vision-language model for segmentation of pneumothorax. *Journal of Imaging Informatics in Medicine*, 37(4):1652–1663, 2024.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, pp. 12. Munich, Germany, 2015.

Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 552–561. Springer, 2020.

Shumeng Li, Jian Zhang, Lei Qi, Luping Zhou, Yinghuan Shi, and Yang Gao. Diversity-enhanced collaborative mamba for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Yueyan Li, Chenggong Zhao, Zeyuan Zang, Caixia Yuan, and Xiaojie Wang. Reading images like texts: Sequential image understanding in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 134066–134104, 2026.

Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE Trans. Med. Imaging.*, 43(1):96–107, 2023.

Yiqun Lin, Huifeng Yao, Zezhong Li, Guoyan Zheng, and Xiaomeng Li. Calibrating label distribution for class-imbalanced barely-supervised knee segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 109–118. Springer, 2022.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. pp. 8801–8809, 2021a.

Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *Medical Imaging with Deep Learning*, 2021b.

Zihe Lv, Zhengda Wu, and Jinghua Zhu. Clustering-guided contrastive prototype learning: Towards semi-supervised medical image segmentation. *Pattern Recognition*, 171:112321, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.112321. URL `https://www.sciencedirect.com/science/article/pii/S0031320325009823`.

Chao Ma and Ziyang Wang. Semi-mamba-unet: Pixel-level contrastive and cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. *Knowledge-Based Systems*, 300:112203, 2024.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyan Huang, et al. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the flare22 challenge. *The Lancet Digital Health*, 6(11):e815–e826, 2024.

Juzheng Miao, Si-Ping Zhou, Guang-Quan Zhou, Kai-Ni Wang, Meng Yang, Shoujun Zhou, and Yang Chen. Sc-ssl: Self-correcting collaborative and contrastive co-training model for semi-supervised medical image segmentation. *IEEE Trans. Med. Imaging.*, 43(4):1347–1364, 2023.

Wenbo Qi, Jiafei Wu, and Shing-Chow Chan. Gradient-aware for class-imbalanced semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pp. 473–490. Springer, 2024.

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

Dandan Shan, Zihan Li, Yunxiang Li, Qingde Li, Jie Tian, and Qingqi Hong. Stpnet: Scale-aware text prompt network for medical image segmentation. *IEEE Trans. Image Process.*, 2025.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Rui Sun, Huayu Mai, Wangkai Li, and Tianzhu Zhang. Towards unbiased learning in semi-supervised semantic segmentation. In *International Conference on Learning Representations*, volume 2025, pp. 63245–63276, 2025.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Haonan Wang and Xiaomeng Li. Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 582–591. Springer, 2023.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, pp. 3876, 2022.

Song Wu, Xiaoyu Wei, Xinyue Chen, Yazhou Ren, Jing He, and Xiaorong Pu. Cross-view mutual learning for semi-supervised medical image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9253–9261, 2024.

Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 34–43. Springer, 2022.

Feilong Xu, Feiyang Yang, Xiongfei Li, and Xiaoli Zhang. A unified loss for handling inter-class and intra-class imbalance in medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8842–8850, 2025a.

Haiming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26007–26020. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a70ee7ea485e4fd36abbfc4adf591c28-Paper-Conference.pdf.

Wenbo Xu, Huaxi Huang, Yongshun Gong, Litao Yu, Qiang Wu, and Jian Zhang. Hierarchical multi-prototype discrimination: Boosting support-query matching for few-shot segmentation. *IEEE Transactions on Multimedia*, 2025b.

Le Yi, Wei Huang, Lei Zhang, Kefu Zhao, Yan Wang, and Zizhou Wang. Dualfete: revisiting teacher-student interactions from a feedback perspective for semi-supervised medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27852 –27860, 2026.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, pp. 605–613. Springer, 2019.

Yuchen Yuan, Xi Wang, Xikai Yang, and Pheng-Ann Heng. Effective semi-supervised medical image segmentation with probabilistic representations and prototype learning. *IEEE Transactions on Medical Imaging*, 44(3):1181–1193, 2025. doi: 10.1109/TMI.2024.3484166.

Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International conference on medical image computing and computer-assisted intervention*, pp. 408–416. Springer, 2017.

Haochen Zhao, Hui Meng, Deqian Yang, Xiaozheng Xie, Xiaoze Wu, Qingfeng Li, and Jianwei Niu. Guidednet: semi-supervised multi-organ segmentation via labeled data guide unlabeled data. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 886–895, 2024.

6 APPENDIX

6.1 ADDITIONAL DATASET

We further evaluate SageProto on AMOS (Ji et al., 2022), which includes 300 CT volumes and 15 anatomical classes: liver, spleen, pancreas (Pan), right kidney (R.Kid), left kidney (L.Kid), stomach (Sto), gallbladder (Gal), esophagus (Eso), aorta, inferior vena cava (IVC), right adrenal gland (RAG), left adrenal gland (LAG), duodenum (Duo), bladder (Bl), and prostate/uterus (P/U). We adopt the same 6:2:2 data split as (Zhao et al., 2024).

6.2 EXPERIMENTAL CONFIGURATION

SageProto follows the Mean Teacher framework with a U-Mamba student and an EMA teacher. Axial CT slices are normalized to $[0, 1]$. FLARE22 and AMOS are resized to 512×512, while Synapse retains its native resolution and is zero-padded to multiples of 32. Training uses SGD with momentum 0.9 and weight decay 10^{-4}. Synapse, FLARE22, and AMOS are trained for 30k, 80k, and 40k iterations with learning rates of 10^{-2}, 5×10^{-4}, and 10^{-3}, respectively. A polynomial learning-rate schedule is applied after 500 warm-up iterations. SageProto components use different activation points and ramp-up schedules across datasets. The teacher is updated by EMA with decay rates for each dataset. The pseudo-label confidence threshold is fixed at 0.75 for Synapse and annealed from 0.82 to 0.76 and 0.70 for FLARE22 and AMOS, respectively. Gradient clipping is disabled for Synapse and set to 12 for FLARE22 and AMOS. Evaluation is performed slice-wise on Synapse and volume-wise on FLARE22 and AMOS. Other hyperparameters are listed in Table 3.

Table 3: Implementation details and hyperparameter settings for SageProto.

Hyperparameter	**Synapse**	**FLARE22**	**AMOS**
Anchor dimension d	256	256	256
Encoder stages L	6	6	6
α_{sct} ceiling	0.08	0.06	0.08
Rectification steps K	4	4	4
Rectification strength λ_c		Class-wise learnable	
Decoder stages J	5	5	5
Prototype weight η	0.06	0.06	0.06
Optimizer	SGD	SGD	SGD
Momentum	0.9	0.9	0.9
Weight decay	10^{-4}	10^{-4}	10^{-4}
Initial learning rate	10^{-2}	$5{\times}10^{-4}$	10^{-3}
Training iterations	30k	80k	40k
LR schedule		Polynomial decay	
Warm-up		500 iterations	
EMA decay	0.99	$0.997{\rightarrow}0.9995$	$0.995{\rightarrow}0.999$
Gradient clipping	None	12	12
Confidence threshold τ	0.75	$0.82{\rightarrow}0.76$	$0.82{\rightarrow}0.70$

6.3 SAGEPROTO ALGORITHM

Algorithm 1 summarizes the procedure of SageProto. Given labeled and unlabeled images, anatomical descriptions and modality information are encoded into class-specific semantic anchors $\{T_c\}_{c=1}^{C}$ shared by the student and EMA teacher. Semantic-Anchor Visual Modulation introduces the anchors through class-dependent affine transformations of the input and hierarchical injection into encoder features, incorporating class semantics across multiple scales. The semantic-aware features are matched with projected anchors $W_t T_c$ to obtain class-wise semantic weights and initialize $\widetilde{P}_c^{(0)}$. Semantic-Anchor Prototype Estimation constructs the visual centroid A_c from ground-truth masks for labeled samples and EMA teacher predictions for unlabeled samples. At rectification step k, the centroid-centered feature–prototype dissimilarity $d_c^{(k)}(p)$ suppresses inconsistent responses and

Algorithm 1 Training procedure of SageProto.

Require: Labeled batch $\mathcal{B}_l$, unlabeled batch $\mathcal{B}_u$, student f_{θ_s}, EMA teacher f_{θ_t}.
Ensure: Updated student and teacher parameters.

Semantic-Anchor Visual Modulation

Encode class descriptions and modality information into semantic anchors $\{T_c\}_{c=1}^{C}$.
Modulate the input and hierarchical encoder features in both branches to obtain semantic-aware visual representations.

Semantic-Anchor Prototype Estimation

Match semantic-aware features with class anchors to estimate $\widetilde{P}_c^{(0)}$.
Compute class-specific visual centroids A_c from labeled masks or unlabeled class activations.
for $k = 0$ **to** $K - 1$ **do**
    Update semantic weights using centroid-centered dissimilarity and obtain $\widetilde{P}_c^{(k+1)}$.
end for

Query-Adaptive Prototype Interaction

Initialize $R_c^{\mathrm{QA},(0)} = \widetilde{P}_c^{(K)}$.
for $j = 1$ **to** J **do**
    Generate prototype-conditioned textual and visual queries from $R_c^{\mathrm{QA},(j-1)}$.
    Retrieve class-specific decoder information and update $R_c^{\mathrm{QA},(j)}$ through adaptive fusion.
end for

Prototype Regularization

Regularize the learned prototypes using student-teacher consistency, semantic correspondence, and inter-class geometry to obtain $\mathcal{L}_{\mathrm{PR}}$.
Compute supervised loss on $\mathcal{B}_l$ and confidence-gated unsupervised loss on $\mathcal{B}_u$.
Optimize $\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \mu(t)\mathcal{L}_{\mathrm{unsup}} + \eta\mathcal{L}_{\mathrm{PR}}$.
Update θ_t by EMA of θ_s.

Table 4: Comparison with state-of-the-art semi-supervised methods on AMOS with 10% labeled data.

Method	Venue	with 10% Labeled Data on the AMOS Dataset															Average	
		Spleen	R.Kid.	L.Kid.	Liver	Sto.	Bla.	Gal.	Eso.	Aorta	IVC	Pan.	RAG	LAG	Duo.	P/U	DSC (%)	Jac. (%)
MT (Tarvainen & Valpola, 2017)	NeurIPS	84.27	85.98	84.42	89.26	56.87	51.22	42.17	50.71	85.57	70.13	48.91	43.92	30.46	48.04	39.46	61.44	51.00
DTC (Luo et al., 2021a)	AAAI	87.48	85.84	85.20	89.81	50.49	64.49	38.62	50.56	85.83	72.72	43.44	42.53	37.02	47.36	39.19	60.81	50.84
CPS (Chen et al., 2021)	CVPR	83.25	87.97	86.30	88.52	55.52	60.53	37.95	52.37	85.36	71.71	50.11	46.33	37.48	46.05	60.33	63.52	51.82
MagicNet (Chen et al., 2023)	CVPR	83.52	87.24	**88.39**	88.99	61.20	62.47	**56.69**	55.68	83.69	74.83	54.11	46.87	43.16	51.18	54.62	65.31	54.89
GA (Qi et al., 2024)	ECCV	70.96	83.12	80.04	87.21	68.81	68.87	50.35	64.97	75.19	65.12	65.67	50.34	52.24	56.58	63.44	66.86	53.02
GuidedNet (Zhao et al., 2024)	ACMMM	**87.50**	87.25	85.86	89.08	66.44	70.65	48.95	59.87	**87.93**	76.32	58.38	54.11	43.40	55.55	67.68	69.19	56.97
DiffMatch (Sun et al., 2025)	ICLR	73.59	85.02	81.95	88.70	70.67	70.86	51.20	66.71	77.77	69.00	66.58	53.38	55.14	58.12	65.87	68.97	55.58
ConformalSAM (Chen et al., 2025)	ICCV	76.27	86.07	83.87	89.35	73.03	73.36	54.44	67.71	80.42	72.38	68.42	59.61	54.98	58.87	67.69	71.10	58.04
DualFete (Yi et al., 2026)	AAAI	75.98	85.54	83.10	89.18	72.26	72.35	53.50	66.67	79.78	71.50	68.11	55.82	**56.17**	58.93	66.98	70.39	57.21
SemiGDA (Huang et al., 2026)	CVPR	77.45	86.15	84.07	87.33	75.07	73.20	54.88	68.07	82.35	76.82	69.50	**60.18**	55.17	**59.95**	69.54	71.98	59.08
SageProto (Ours)	–	81.85	**88.26**	86.18	**90.94**	**75.91**	**76.80**	55.53	**70.02**	85.25	**77.17**	**70.75**	58.31	56.01	59.45	**71.31**	**73.58**	**61.29**
Δ vs. SemiGDA		**↑4.40%**	**↑2.11%**	**↑2.11%**	**↑3.61%**	**↑0.84%**	**↑3.60%**	**↑0.65%**	**↑1.95%**	**↑2.90%**	**↑0.35%**	**↑1.25%**	**↓1.87%**	**↑0.84%**	**↓0.50%**	**↑1.77%**	**↑1.60%**	**↑2.21%**

updates $\widetilde{P}_c^{(k+1)}$. After K iterations, $\widetilde{P}_c^{(K)}$ initializes Query-Adaptive Prototype Interaction. At each decoder layer, the prototype generates textual and visual queries, retrieves relevant features, and fuses them into $R_c^{\mathrm{QA},(j)}$, which is passed to the next layer. Prototype Regularization constrains student-teacher prototype consistency, prototype–anchor semantics, and inter-class geometry. The student is optimized with supervised segmentation, confidence-gated unsupervised consistency, and prototype regularization, while the teacher is updated by exponential moving average.

6.4 Experimental Analysis

Comparison with State-of-the-Art Methods. Table 4 compares SageProto with existing semi-supervised methods on AMOS under 10% labeled supervision. SageProto reaches 73.58% DSC and 61.29% Jaccard, exceeding SemiGDA by 2.22 and 3.74 points, respectively. Class-wise gains include 5.68 points on spleen, 4.13 on liver, 4.92 on bladder, 2.86 on esophagus, 3.52 on aorta, 1.80 on pancreas, and 2.55 on prostate/uterus. SageProto attains the best performance on several challenging structures, including stomach, bladder, esophagus, IVC, pancreas, and prostate/uterus. Despite decreases of 3.11 and 0.83 points on the right adrenal gland and duodenum, SageProto improves 13 of the 15 classes. The gains are concentrated on difficult classes, indicating that class-specific semantic references throughout representation learning, prototype construction, and decoding reduce class-wise learning disparity under limited supervision.

Comparison with State-of-the-Art Methods under Increased Supervision. Tables 5, 6, and 7 report results under increased labeled supervision. On Synapse with 20% labeled data, SageProto reaches 72.94% DSC and 5.36 mm ASD, exceeding DCMamba by 5.67 DSC points and reducing ASD by 3.92 mm. Gallbladder, aorta, and pancreas gain 22.93, 12.06, and 9.42 Dice points, respectively, although liver decreases. On FLARE22 with 50% labeled data, SageProto achieves 86.21% DSC and 78.61% Jaccard, exceeding GuidedNet by 3.09 and 6.14 points. Larger gains occur on the right and left adrenal glands, stomach, and esophagus, while gallbladder and duodenum improve less. On AMOS with 50% labeled data, SageProto exceeds GuidedNet by 2.49 DSC points and 3.14 Jaccard points, with larger gains on pancreas, left adrenal gland, prostate/uterus, and esophagus. Across the three datasets, SageProto retains its advantage as labeled supervision increases, with larger improvements on difficult classes and smaller changes on well-learned classes. The class-wise gain pattern matches the role of semantic references throughout representation learning, prototype construction, and decoding.

Table 5: Comparison with state-of-the-art semi-supervised methods on Synapse with 20% labeled data.

Method	Venue	20% Labeled Data on Synapse								Average	
		R.Kidney	L.Kidney	Liver	Spleen	Stomach	Aorta	Gallbladder	Pancreas	DSC (%) ↑	ASD (mm) ↓
MT (Tarvainen & Valpola, 2017)	NeurIPS	44.08	40.02	72.43	55.56	38.56	67.71	29.11	14.85	45.29	34.61
FixMatch (Sohn et al., 2020)	NeurIPS	57.87	64.47	80.45	75.42	49.72	79.72	45.03	30.10	60.35	17.61
CPS (Chen et al., 2021)	CVPR	42.46	45.00	75.41	55.80	37.30	68.42	32.34	15.59	46.54	28.51
SS-Net (Wu et al., 2022)	MICCAI	51.26	50.78	76.76	66.85	41.83	72.71	35.20	13.13	51.07	24.68
BCP (Bai et al., 2023)	CVPR	72.19	62.74	79.38	75.25	52.73	77.28	25.06	37.36	60.25	25.50
IC (Huang et al., 2024)	TMI	71.11	67.48	88.02	79.19	39.54	74.50	35.82	10.67	58.29	15.09
CML (Wu et al., 2024)	ACMMM	51.97	57.70	67.74	61.93	28.80	71.70	9.85	4.21	44.24	46.17
Semi-Mamba-UNet (Ma & Wang, 2024)	KBS	71.64	65.67	82.87	78.20	54.21	70.54	25.74	35.79	60.58	10.99
DCMamba (Li et al., 2025)	TMI	80.80	78.28	**91.60**	81.00	62.70	75.17	36.09	32.52	67.27	9.28
SageProto (Ours)	–	**84.38**	**84.78**	78.17	**81.75**	**66.28**	**87.23**	**59.02**	**41.94**	**72.94**	**5.36**
Δ vs. DCMamba		**↑3.58%**	**↑6.50%**	**↓13.43%**	**↑0.75%**	**↑3.58%**	**↑12.06%**	**↑22.93%**	**↑9.42%**	**↑5.67%**	**↓3.92%**

Table 6: Comparison with state-of-the-art semi-supervised methods on FLARE22 with 50% labeled data.

Method	Venue	50% Labeled Data on FLARE22													Average	
		Spleen	R.Kid.	L.Kid.	Liver	Sto.	Gal.	Eso.	Aorta	IVC	Pan.	RAG	LAG	Duo.	DSC (%) ↑	Jac. (%) ↑
DAN (Zhang et al., 2017)	MICCAI	89.74	93.24	93.63	96.50	62.04	70.07	67.20	90.76	80.09	61.71	58.04	43.47	66.60	74.86	63.73
MT (Tarvainen & Valpola, 2017)	NeurIPS	89.64	92.58	93.66	96.97	66.63	69.89	71.68	91.39	82.08	68.65	57.51	62.21	60.96	77.22	65.94
SASSNet (Li et al., 2020)	MICCAI	92.03	93.79	92.47	95.25	66.48	65.57	70.83	90.03	79.94	63.61	59.98	62.78	60.14	76.38	65.20
DTC (Luo et al., 2021a)	AAAI	91.33	93.57	94.46	96.47	65.94	75.80	68.53	92.52	83.77	64.88	68.87	61.35	65.58	78.70	67.64
CPS (Chen et al., 2021)	CVPR	91.23	93.70	93.38	96.77	72.63	78.81	72.37	92.35	83.30	70.34	58.34	61.47	65.48	79.24	67.99
CLD (Lin et al., 2022)	MICCAI	88.54	93.33	91.48	94.27	74.88	76.68	71.03	91.51	83.21	71.51	64.57	63.98	68.15	79.47	67.95
DHC (Wang & Li, 2023)	MICCAI	90.81	92.18	93.39	92.54	76.96	84.60	72.91	91.84	83.25	74.65	64.52	55.88	69.44	80.23	68.90
MagicNet (Chen et al., 2023)	CVPR	88.84	93.06	90.84	96.49	80.33	82.81	75.79	91.72	81.98	69.46	63.40	59.98	67.44	80.16	69.08
GuidedNet (Zhao et al., 2024)	ACMMM	92.10	93.64	93.45	96.09	80.60	85.26	76.09	92.44	84.28	75.73	71.34	68.33	71.15	83.12	72.47
SageProto (Ours)	–	**97.72**	**96.21**	93.84	**97.34**	**88.06**	76.69	**82.33**	**93.80**	**87.30**	**75.92**	**82.24**	**82.41**	66.85	**86.21**	**78.61**
Δ vs. GuidedNet		**↑5.62%**	**↑2.57%**	**↑0.39%**	**↑1.25%**	**↑7.46%**	**↓8.57%**	**↑6.24%**	**↑1.36%**	**↑3.02%**	**↑0.19%**	**↑10.90%**	**↑14.08%**	**↓4.30%**	**↑3.09%**	**↑6.14%**

Table 7: Comparison with state-of-the-art semi-supervised methods on AMOS with 50% labeled data.

Method	Venue	50% Labeled Data on AMOS														Average	
		Spleen	R.Kid.	L.Kid.	Liver	Sto.	Bla.	Gal.	Eso.	Aorta	IVC	Pan.	RAG	LAG	Duo.	P/U	DSC (%) ↑	Jac. (%) ↑
DAN (Zhang et al., 2017)	MICCAI	89.63	88.74	90.08	90.49	55.91	47.44	39.13	58.34	86.71	72.09	54.98	29.57	6.49	50.33	53.04	61.39	52.06
MT (Tarvainen & Valpola, 2017)	NeurIPS	89.83	89.24	90.23	92.08	62.02	63.05	51.63	59.64	89.12	78.11	53.46	45.41	37.35	52.85	40.93	66.17	57.06
UA-MT (Yu et al., 2019)	MICCAI	88.92	88.83	88.93	90.86	58.55	54.86	40.99	58.58	88.49	74.28	51.88	51.31	41.78	54.54	44.73	65.48	55.62
SASSNet (Li et al., 2020)	MICCAI	91.54	89.72	89.61	91.65	53.00	50.43	37.93	60.57	88.50	74.87	46.34	45.62	39.17	52.48	55.92	63.77	54.68
DTC (Luo et al., 2021a)	AAAI	90.68	89.30	88.88	91.25	56.49	67.37	41.88	62.49	89.16	76.50	48.13	47.67	42.91	54.67	54.23	66.93	55.92
CPS (Chen et al., 2021)	CVPR	89.97	89.67	90.25	90.94	61.90	65.03	42.87	62.44	88.77	75.27	52.34	49.96	47.74	45.15	54.76	66.65	56.56
CLD (Lin et al., 2022)	MICCAI	89.34	89.86	89.50	91.23	66.18	66.40	45.78	62.93	88.85	76.97	55.63	54.24	43.79	53.35	58.82	69.09	57.99
DHC (Wang & Li, 2023)	MICCAI	86.62	87.48	85.57	86.68	58.39	67.04	53.75	54.14	87.28	74.38	60.88	51.59	51.03	56.91	58.87	68.60	56.05
MagicNet (Chen et al., 2023)	CVPR	88.59	89.64	90.28	91.69	66.33	61.80	58.83	59.53	86.80	74.39	59.94	52.74	42.35	52.88	57.28	68.94	58.33
GuidedNet (Zhao et al., 2024)	ACMMM	91.21	89.87	90.82	92.32	72.99	74.00	56.72	66.62	89.31	78.41	61.75	54.71	51.10	58.26	65.23	72.94	61.53
SageProto (Ours)	–	83.69	**90.03**	87.17	**92.55**	**76.61**	**77.79**	**58.24**	**72.22**	86.50	77.97	**73.61**	**59.84**	**60.20**	**62.21**	**72.87**	**75.43**	**64.67**
Δ vs. GuidedNet		**↓7.52%**	**↑0.16%**	**↓3.65%**	**↑0.23%**	**↑3.62%**	**↑3.79%**	**↑1.52%**	**↑5.60%**	**↓2.81%**	**↓0.44%**	**↑11.86%**	**↑5.13%**	**↑9.10%**	**↑3.95%**	**↑7.64%**	**↑2.49%**	**↑3.14%**

Quantitative Analysis. Tables 8 and 9 evaluate the contribution of each SageProto component. The baseline relies on visual evidence for representation and prototype learning, leaving poorly learned classes with weaker class representations. SAVM introduces class-specific semantic references into representation learning, improving DSC from 78.11% to 80.47% on FLARE22 and from 63.38% to 65.77% on AMOS. SAPE uses semantic-anchor initialization and iterative rectification, reaching 83.65% and 67.24% DSC, while QAPI retrieves textual and visual features through class-specific prototypes, achieving 82.77% and 66.76%. Combining SAVM and SAPE raises DSC to 84.55% and 70.88%, and adding QAPI increases it to 85.47% and 72.14%. With Prototype Regularization, SageProto reaches 86.21% DSC and 77.04% Jaccard on FLARE22, and 73.58% DSC

Table 8: Ablation study of SageProto on FLARE22 with 10% labeled data. SAVM, SAPE, QAPI, and PR denote Visual Modulation, Prototype Estimation, Prototype Interaction, and Regularization.

Method	with 10% Labeled Data on the FLARE22 Dataset													Average	
	Spleen	R.Kid.	L.Kid.	Liver	Sto.	Gal.	Eso.	Aorta	IVC	Pan.	RAG	LAG	Duo.	DSC (%)	Jac. (%)
Baseline	94.68	91.74	86.61	69.87	87.81	76.36	71.94	75.70	58.31	68.86	80.33	64.83	88.43	78.11	67.34
+ SAVM	95.93	92.60	95.60	55.18	89.33	82.33	82.82	81.14	75.76	80.99	77.05	47.09	90.36	80.47	69.72
+ SAPE	96.03	93.32	95.89	68.57	90.40	85.13	82.38	81.76	77.33	81.16	82.13	61.93	91.41	83.65	73.28
+ QAPI	96.32	94.18	96.24	65.44	90.66	84.75	82.44	80.09	75.41	81.73	81.71	55.55	91.47	82.77	72.39
SAVM + SAPE	96.64	95.14	97.00	71.81	92.12	85.62	82.97	81.76	74.86	82.29	85.02	61.61	92.33	84.55	74.71
SAVM + SAPE + QAPI	97.06	95.65	97.45	73.58	92.81	86.50	82.77	81.96	76.68	81.86	86.40	65.36	93.04	85.47	75.93
SageProto	**97.34**	**96.21**	**97.72**	**75.92**	**93.80**	**87.30**	**82.24**	**82.41**	**76.69**	**82.33**	**88.06**	**66.85**	**93.84**	**86.21**	**77.04**
Δ vs. Baseline	**↑2.66%**	**↑4.47%**	**↑11.11%**	**↑6.05%**	**↑6.00%**	**↑10.94%**	**↑10.30%**	**↑6.71%**	**↑18.38%**	**↑13.47%**	**↑7.73%**	**↑2.02%**	**↑5.41%**	**↑8.10%**	**↑9.70%**

Table 9: Ablation study of SageProto on AMOS with 10% labeled data. SAVM, SAPE, QAPI, and PR denote Visual Modulation, Prototype Estimation, Prototype Interaction, and Regularization.

Method	with 10% Labeled Data on the AMOS Dataset										Average	
	Spleen	R.Kid.	L.Kid.	Liver	Sto.	Gal.	Pan.	RAG	LAG	Duo.	DSC (%)	Jac. (%)
Baseline	71.67	82.88	79.39	83.32	61.90	47.28	63.64	44.12	49.50	55.21	63.38	49.44
+ SAVM	72.79	83.67	78.80	85.27	63.06	45.79	66.01	44.27	55.26	57.58	65.77	52.03
+ SAPE	72.95	84.59	82.04	84.48	66.04	53.52	66.75	48.17	52.66	57.85	67.24	53.41
+ QAPI	74.58	82.00	78.96	85.69	66.30	53.08	66.42	48.44	51.90	57.22	66.76	52.93
SAVM + SAPE	76.72	86.01	84.11	87.67	71.53	56.10	69.29	55.74	56.09	60.11	70.88	57.65
SAVM + SAPE + QAPI	79.42	87.77	85.61	88.26	73.26	55.27	70.50	56.62	55.65	60.28	72.14	59.25
SageProto	**81.85**	**88.25**	**86.18**	**90.94**	**75.91**	**55.53**	**70.75**	**58.31**	**56.01**	**59.45**	**73.58**	**61.29**
Δ vs. Baseline	**↑10.18%**	**↑5.37%**	**↑6.79%**	**↑7.62%**	**↑14.01%**	**↑8.25%**	**↑7.11%**	**↑14.19%**	**↑6.51%**	**↑4.24%**	**↑10.20%**	**↑11.85%**

and 61.29% Jaccard on AMOS. The improvements across these configurations show that each stage contributes to reducing the reliance of difficult classes on weak visual evidence.

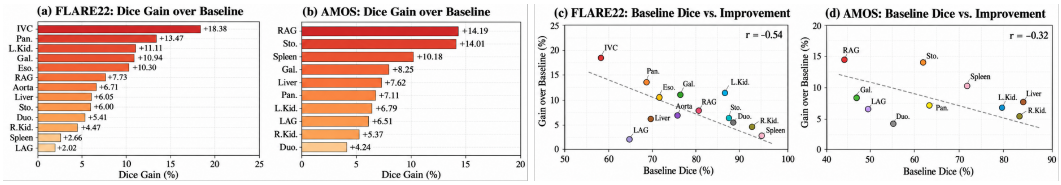
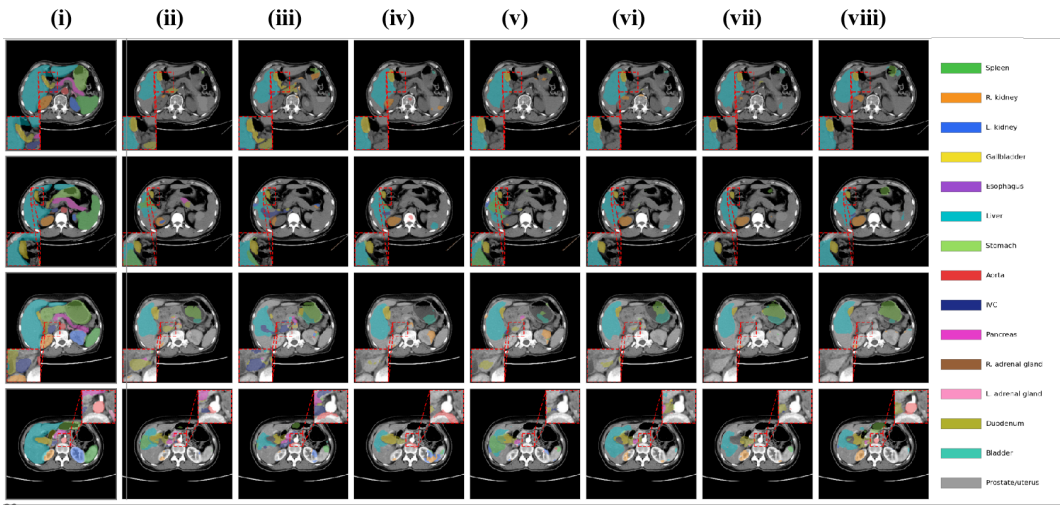


Figure 4: Class-wise analysis of SageProto on FLARE22 and AMOS with 10% labeled data. (a) FLARE22 Dice improvements over the baseline across individual organs, highlighting larger gains on poorly learned classes. (b) AMOS Dice improvements over the baseline, showing a similar class-dependent improvement pattern. (c) FLARE22 baseline Dice versus performance gain, with a negative correlation indicating larger gains for classes with lower baseline performance. (d) AMOS baseline Dice versus performance gain, further showing that weaker classes tend to benefit more from SageProto.

Class-wise Analysis. SageProto mainly improves classes poorly learned by the baseline. On FLARE22, IVC and pancreas gain 18.38 and 13.47 Dice points, while on AMOS, the right adrenal gland and stomach improve by 14.19 and 14.01 points. These classes have relatively low baseline Dice scores, whereas well-learned organs change less. Fig. 4(c,d) shows a negative correlation between baseline Dice and performance gain on FLARE22 ($r = -0.54$) and AMOS ($r = -0.32$), with larger improvements concentrated on weaker classes. Under limited supervision, strong classes form reliable visual representations earlier, while difficult classes remain weakly learned and receive less reliable pseudo-supervision. SageProto introduces class-specific semantic references throughout representation learning, prototype construction, and decoding, reducing the dependence of difficult classes on weak visual evidence. The gains are concentrated on poorly learned classes, narrowing the class-wise learning gap rather than uniformly improving all organs.

Qualitative Analysis. Fig. 5 further illustrates how the individual components affect class-wise segmentation on AMOS and FLARE22. The baseline generally segments large and well-learned organs reliably, but frequently misses small, low-contrast, or weakly learned structures, as highlighted by the enlarged regions. Introducing SAVM, SAPE, or QAPI individually recovers part of these structures, although local omissions and inaccurate boundaries remain. Combining SAVM and SAPE produces more complete class regions, and further incorporating QAPI improves the continu-



(a) Quantitative Results for AMOS.

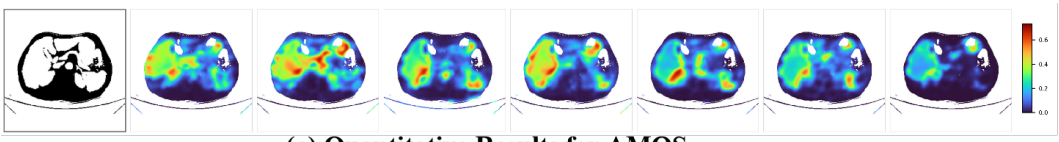


(b) Quantitative Results for Flare22.

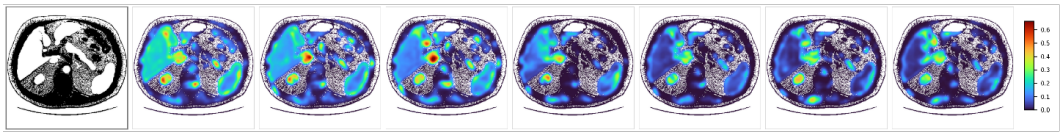
Figure 5: Qualitative comparison of SageProto components on AMOS and FLARE22 with 10% labeled data. (i) Ground Truth, (ii) Baseline, (iii) +SAVM, (iv) +SAPE, (v) +QAPI, (vi) SAVM+SAPE, (vii) SAVM+SAPE+QAPI, and (viii) SageProto. Red dashed boxes highlight challenging regions, with enlarged views showing the recovery of small and poorly learned structures as components are added.

ity of difficult structures across the decoder stages. The full SageProto predictions are consistently closer to the ground truth, particularly for classes that are poorly represented by the baseline. This behavior agrees with the quantitative results on FLARE22, where the largest gains occur on IVC, pancreas, gallbladder, and esophagus, with improvements of 18.38, 13.47, 10.94, and 10.30 Dice points, respectively, whereas already well-learned classes such as spleen show smaller changes. The visual results therefore reinforce the class-wise analysis: semantic anchors improve classes whose visual representations remain weak under limited supervision, reducing the disparity between poorly learned and well-learned organs.

Predictive Uncertainty Evolution. Fig. 6 shows predictive uncertainty during training. At early stages, high-entropy responses spread across the abdominal region, covering organ interiors and surrounding tissues, while class predictions remain unstable. As training proceeds, these responses weaken and become more localized. Mean predictive entropy decreases from 0.069 to 0.026 on AMOS and from 0.053 to 0.027 on FLARE22, corresponding to reductions of 62% and 50%, re-



(a) Quantitative Results for AMOS.



(a) Quantitative Results for Flare22.

Figure 6: Predictive uncertainty evolution of SageProto on AMOS, and FLARE22 with 10% labeled data. As training progresses, high-entropy responses gradually contract from diffuse abdominal regions toward ambiguous anatomical boundaries, while organ interiors become increasingly certain.

spectively. The reduction is spatially non-uniform: uncertainty decreases earlier within confidently predicted organ interiors, while stronger responses persist around small structures and ambiguous anatomical interfaces. At later stages, most interior regions exhibit low entropy, with residual uncertainty concentrated near organ boundaries, adjacent structures, and low-contrast regions. The final maps are dominated by anatomically ambiguous regions rather than diffuse interior responses.