FM-ABS: Promptable Foundation Model Drives Active Barely Supervised Learning for 3D Medical Image Segmentation

Zhe Xu[1,2], Cheng Chen[2], Donghuan Lu[3], Jinghan Sun[5], Dong Wei[3], Yefeng Zheng[3], Quanzheng Li[2,4(✉)], and Raymond Kai-yu Tong[1(✉)]

[1] Department of Biomedical Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
jackxz@link.cuhk.edu.hk; kytong@cuhk.edu.hk
[2] Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
li.quanzheng@mgh.harvard.edu
[3] Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China
[4] Data Science Office, Massachusetts General Brigham, Boston, MA, 02116, USA
[5] Xiamen University, Xiamen, China

Abstract. Semi-supervised learning (SSL) has significantly advanced 3D medical image segmentation by effectively reducing the need for laborious dense labeling from radiologists. Traditionally focused on *model-centric* advancements, we anticipate that the SSL landscape will shift due to the emergence of open-source generalist foundation models, e.g., Segment Anything Model (SAM). These generalists have shown remarkable zero-shot segmentation capabilities with manual prompts, allowing a promising *data-centric* perspective for future SSL, particularly in pseudo and expert labeling strategies for enhancing the data pool. To this end, we propose the Foundation Model-driven Active Barely Supervised (FM-ABS) learning paradigm for developing customized 3D specialist segmentation models with shoestring annotation budgets, i.e., merely labeling three slices per scan. Specifically, building upon the basic mean-teacher framework, FM-ABS accounts for the intrinsic characteristics of 3D imaging and modernizes the SSL paradigm with two key data-centric designs: (i) specialist-generalist collaboration where the in-training specialist model delivers class-specific prompts to interact with the frozen class-agnostic generalist model across multiple views to acquire noisy-yet-effective pseudo labels, and (ii) expert-model collaboration that advocates active cross-labeling with notably low annotation efforts to progressively provide the specialist model with informative and efficient supervision in a human-in-the-loop manner, which benefits the automatic object-specific prompt generation in turn. Extensive experiments on two benchmark datasets show the promising results of our approach over recent SSL methods under extremely limited (barely) labeling budgets.

Keywords: Barely Supervised · Foundation Model · Cross Labeling.

1 Introduction

Medical image segmentation plays a crucial role in many image-guided therapies, and deep learning (DL)-based methods have greatly advanced automatic organ or tumor segmentation [9, 11]. Yet, these successes typically hinge on the availability of extensive densely labeled data, which is generally expertise-demanding, expensive and laborious to obtain. Since unlabeled data is often abundant in practice, semi-supervised learning (SSL) presents an attractive solution by effectively utilizing both labeled and unlabeled data.

Typically, the SSL paradigm involves randomly selecting samples for radiologists to densely annotate based on the available budget, and then advanced SSL algorithms utilize this mix of densely labeled and unlabeled data for training. Considerable efforts on SSL have been made, with popular tracks including self-training [1, 6, 20, 7, 23], consistency regularization [15, 24, 26, 12, 25, 22] and adversarial training [28]. These advancements predominantly follow a *model-centric* paradigm, with performance inherently depending on the knowledge transfer from labeled to unlabeled data. As such, most methods still appreciate optimistic budgets for high-quality densely labeled data, struggling to achieve satisfactory outcomes when faced with extremely limited (barely) labeling budgets [4].

Recently, generalist segmentation foundation models, featured by the Segment Anything Model (SAM) [10, 27], have shown impressive zero-shot segmentation capabilities with manual prompts. However, recent work reveals that SAM often struggles to deliver fine-grained results in various scenarios, including medical imaging [8], where manual prompting is also notably laborious. Thus, the need for fully automatic specialist models remains important. To utilize SAM in medical images more effectively, Ma et al. finetuned SAM on substantial labeled medical images [13], while Chen et al. [5] finetuned and extended the 2D SAM to 3D with adapters. Despite the promising results, these approaches are label- or computation-intensive and offer limited customization when the specialist model is expected to be architecturally and functionally tailored to practical requirements. Thus, rather than finetuning them, we expect these generalist models to act as catalysts for new *data-centric* opportunities in the future SSL landscape, such as (i) automatically scaling up the labeled pool for comprehensive supervised signals; and (ii) optimizing the use of precious annotation budgets.

To this end, we propose the Foundation Model-driven Active Barely Supervised (FM-ABS) learning paradigm, as depicted in Fig. 1, for developing customized 3D specialist segmentation models with scarce annotation budgets. Specifically, building upon the basic mean-teacher framework with consistency regularization, FM-ABS accounts for the intrinsic characteristics of 3D imaging and further modernizes the SSL paradigm with two *data-centric* designs: (i) **Specialist-generalist collaboration** (Fig. 1 (a)), where the in-training specialist model delivers class-specific prompts (i.e., boxes and points) to interact with the frozen class-agnostic generalist model to acquire complementary pseudo supervision. Due to the inherent lack of depth perception of the 2D generalist and the ambiguity between object and background in medical images, the generalist-based labels are very noisy. Thus, we propose generating pseudo labels across the

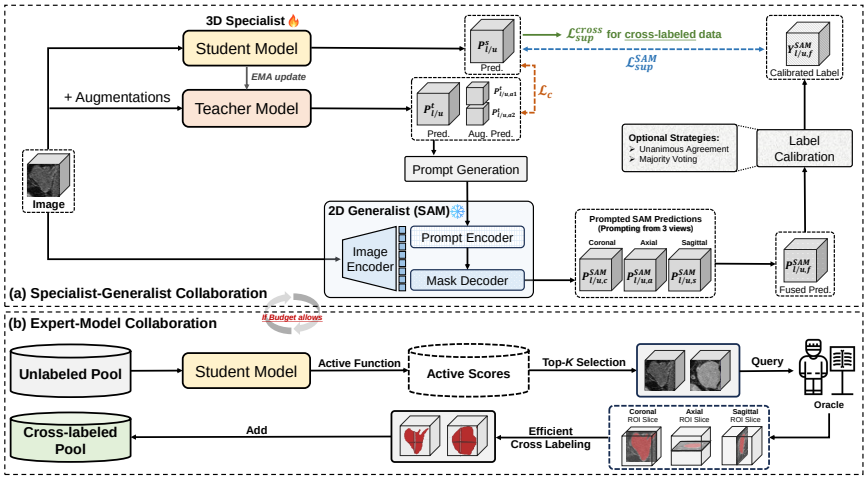

Fig. 1. Overview of our Foundation Model-driven Active Barely Supervised (FM-ABS) learning paradigm, consisting of (a) the specialist-generalist collaboration scheme (Sec. 2.2) and (b) the expert-model collaboration scheme (Sec. 2.3).

three views (axial, sagittal and coronal) and obtaining final labels through consensus mechanisms. Equipped with the noise-tolerant loss, such extended noisy supervision proves effective in the scarce supervision regime. (ii) **Expert-model collaboration**, which advocates active multi-view cross-labeling (Fig. 1 (b)) for 3D images with notably low annotation efforts (merely labeling three orthogonal slices per scan). This labeling strategy considers the intrinsic anatomical similarity among adjacent 2D slices and the informative disparities across views. It aims to progressively provide the local specialist model with informative and efficient supervision in a human-in-the-loop manner to meet the model-informed needs. This collaboration, in turn, facilitates automatic object-specific prompt generation. Our method is evaluated on two benchmark datasets and shows promising performance under the extremely limited (barely) annotation budget.

2 Method

2.1 Preliminaries

Traditional SSL typically involves randomly preselecting an M-sample subset from dataset $\mathcal{D}$ for dense labeling, after which both the labeled subset $\mathcal{D}^l$ (M samples) and the unlabeled subset $\mathcal{D}^u$ (N samples) are used for training. Here, we explore a new active barely supervised learning (BSL) setting, emphasizing more efficient use of extremely limited annotation budgets. Unlike traditional SSL, our active BSL incorporates a human-in-the-loop paradigm to select model-informed informative samples for labeling during the training process, and employs an efficient cross-labeling strategy (elaborated in Sec. 2.3). Specifically, at the r-th

round of active learning, we have a cross-labeled subset $\mathcal{D}_r^{cl} = \{(X_i^{cl}, Y_i^{cl})\}_{i=1}^{N_r}$ with N_r cross-labeled scans, and the remainder as the unlabeled subset $\mathcal{D}_r^u = \{X_i^u\}_{i=N_r+1}^{N_r+M_r}$ with M_r unlabeled scans. $X_i^{cl}, X_i^u \in \mathbb{R}^{H\times W\times D}$ denote the scans with height H, width W and depth D, and $Y_i^{cl} \in \{0,1\}^{H\times W\times D}$ denotes the cross-label of X_i^{cl} (we focus on binary segmentation). Before training, we initiate the process by randomly selecting a small subset of scans for labeling, thus constructing $\mathcal{D}_0^{cl}$ as the starting point. Our goal is to learn segmentation with active cross-labeled data and unlabeled data by optimizing the following loss:

$$\mathcal{L} = \mathcal{L}_{sup}^{cross}(\mathcal{D}_r^{cl}) + \mathcal{L}_{aux}(\mathcal{D}_r^{cl}, \mathcal{D}_r^u), \tag{1}$$

where $\mathcal{L}_{sup}^{cross}$ and $\mathcal{L}_{aux}$ denote the supervised loss from $\mathcal{D}_r^{cl}$ and auxiliary guidance from all data, respectively. We adopt the partial cross-entropy loss on the labeled voxels for $\mathcal{L}_{sup}^{cross}$. As shown in Fig. 1, our framework is built upon the popular mean-teacher SSL model with two data-centric strategies: the specialist-generalist collaboration (Sec. 2.2) and the expert-model collaboration (Sec. 2.3).

2.2 Specialist-Generalist Collaboration

Mean-Teacher Model. The mean-teacher (MT) model is a standard SSL design powered by consistency regularization. Specifically, the trainable student f_θ^s is optimized by standard back-propagation, while the teacher model $f_{\tilde{\theta}}^t$ updates through an exponential moving average (EMA) of past and current weights [15]. Denoting θ as the student's weights and $\tilde{\theta}$ as the teacher's weights, $\tilde{\theta}$ is updated by $\tilde{\theta}_t = \alpha\tilde{\theta}_{t-1} + (1-\alpha)\theta_t$ at iteration t, where α is the EMA coefficient and empirically set to 0.99 [26]. We also inherit the consistency loss $\mathcal{L}_c$ [26] as part of $\mathcal{L}_{aux}$, formulated as: $\mathcal{L}_c = \frac{1}{J}\cdot\sum_j d\left(f_{\tilde{\theta}}^t(X+\xi_j), f_\theta^s(X)\right)$, where mean absolute error is used as the distance function $d(\cdot,\cdot)$; ξ_j indicates the types of perturbation, wherein Gaussian noise and random contrast adjustments are utilized here. We adopt this stability constraint because (i) it helps regularize model behavior and enhance generalizability [16, 29], alleviating overfitting caused by limited supervision, and (ii) it encourages the model to focus on the intrinsic data structure, reducing the impact of confirmation bias induced by noisy generalist-based pseudo labels as discussed later.

Generalist-based Multi-view Pseudo Label Learning. Cross-labeling enhances annotation efficiency but introduces significant challenges for traditional SSL due to its sparse supervision. In response, we leverage pre-trained 2D generalist models, known for their robust zero-shot capabilities, to assist label augmentation. In this synergy, the specialist model automatically generates prompts for the generalist model, which, in turn, provides auxiliary supervision for the specialist model through its responses to these prompts. The generalist model (e.g., SAM trained on over one billion images) features an architecture comprising an image encoder, a prompt encoder and a mask decoder. It supports both sparse (points and boxes) and dense (grids or masks) prompt formats, allowing flexible and class-agnostic segmentation. The essence of collaboration hinges on

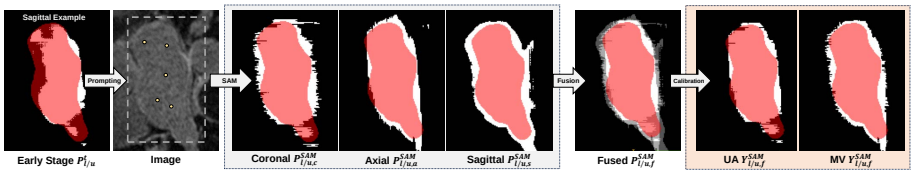

Fig. 2. Workflow of generalist-based label generation for LA MRI. The exemplar sagittal view of $P^t_{l/u}$ is from the early-stage model. Red color represents the ground truth.

(i) automatically crafting effective prompts from the in-training specialist model, (ii) improving the quality of generalist-based pseudo labels, and (iii) enabling the specialist to effectively assimilate knowledge from these noisy labels.

Prompt Generation and Multi-view Pseudo Labeling. As depicted in Fig. 1 (a), the image X is fed into the self-ensembling teacher model to obtain its mask prediction P^t. Acknowledging the 2D generalist's difficulties in providing precise pseudo labels and its lack of 3D perception, we inherit the multi-view labeling spirit that generates object-specific 2D bounding boxes and randomly sampled points from P^t across the three views (c for coronal, a for axial, and s for sagittal) to sparsely prompt the frozen 2D generalist slice-by-slice, thereby acquiring three 3D pseudo labels (Y_c^{SAM}, Y_a^{SAM}, and Y_s^{SAM}). Given the expected inaccuracies in P^t due to limited supervision, we apply a modest random expansion (0-10 pixels) to the prompt boxes. As observed in Fig. 2, the pseudo masks via prompting from the three views are noisy and significantly differ, yet they overlap in certain areas. Thus, we fuse the view-specific masks and introduce two intuitive strategies for label calibration: unanimous agreement (UA), considering labels only when there is consensus across all views; and majority voting (MV), calibrating labels based on the most frequent prediction across views. As such, we obtain the calibrated final label Y_f^{SAM} as exampled in Fig. 2. Note that due to the 2D generalist model's low inference efficiency for 3D scans, we strategically update Y_f^{SAM} in a few task-specific rounds, balancing its utility with computational demands.

Noise-Tolerant Collaborative Learning. Considering the inevitably noisy nature of Y_f^{SAM}, we further introduce a noise-tolerant Dice loss [18] to alleviate the adverse effects of label noise, formulated as:

$$\mathcal{L}_{sup}^{SAM} = \frac{\sum_v |P_v^s - Y_{f,v}^{SAM}|^\gamma}{\sum_v \left(P_v^s\right)^2 + \sum_v \left(Y_{f,v}^{SAM}\right)^2 + \epsilon}, \tag{2}$$

where P_v^s is the predicted probabilities of voxel v from the student model and Y_f^{SAM} is converted to one-hot representation. We set $\gamma = 1.5$ and $\epsilon = 10^{-5}$ due to demonstrated noise-robustness [18]. Note that when $\gamma = 2$, this loss degrades into the typical Dice loss. Such generalist-based labels can provide effective early supervision, yet, overly relying on these noisy labels in the later training stages can mislead the model due to the memorization effect [2]. As for the consistency loss $\mathcal{L}_c$, the model struggles to perceive the object at the early stage, rendering

this stability constraint relatively meaningless. Thus, we introduce two time-dependent trade-off weights, λ_{sam} and λ_c, to modulate the importance of the two losses during the collaborative training. Specifically, $\lambda_{sam}(t) = 0.5 \cdot [1 - e^{-5\left(1-\frac{t}{t_{max}}\right)^2}]$ and $\lambda_c(t) = 0.1 \cdot e^{-5\left(1-\frac{t}{t_{max}}\right)^2}$ [26], where t and $t_{\max}$ denote the current and the maximal iterations, respectively. As such, $\mathcal{L}_{aux}$ is formulated as:

$$\mathcal{L}_{aux}(\mathcal{D}_r^{cl}, \mathcal{D}_r^u) = \lambda_{sam}\mathcal{L}_{sup}^{SAM}(\mathcal{D}_r^{cl}, \mathcal{D}_r^u) + \lambda_c\mathcal{L}_c(\mathcal{D}_r^{cl}, \mathcal{D}_r^u). \tag{3}$$

2.3 Expert-Model Efficient Collaboration

Active Selection. Moving beyond the traditional random selection, FM-ABS embraces multi-round model-informed active selection during training. At each round, we query the unlabeled pool based on model-informed scores for annotation and grow the cross-labeled set, repeating until the budget is exhausted. We kick off the process with a random initial selection using 20% of budget to warm up the model. Our FM-ABS is equipped with a family of efficient active functions, including (i) least confidence (LC) sampling for the top-K cases with minimal average value over the bottom quartile of prediction confidences (probabilities), (ii) classical highest entropy (HE) sampling for top-K cases with maximal mean predictive entropy, and (iii) highest entropy ratio (HER) sampling for the top-K case with the predominant share of high-entropy predictions beyond a time-dependent threshold. K is set to 2 here. More specifically for (ii) and (iii), we denote p_v^s as the predicted probability of the student model at a voxel v, its normalized entropy ne_v is computed by $ne_v = -\sum_{c\in C} p_{c,v}^s \log\left(p_{c,v}^s\right) / \log(|C|) \in [0, 1]$, where $c \in C$ is the semantic label. High entropy indicates a high level of uncertainty to some extent. For HE, we compute the average entropy across all voxels to determine the active score. In HER, we count the voxels with $ne_v > \beta$ (β being an empirical threshold from 0.5 to 0.75 via a Gaussian ramp-up function) and then compute their ratio to the total voxel count, offering a clearer distribution of high uncertainty across the image compared to HE.

Cross-Labeling Strategy. Densely labeling all slices in selected scans, as used in traditional SSL, is budget-intensive and reduces expert-model collaboration's efficiency. Yet, weak labels (e.g., bounding boxes or scribbles) often result in only rough boundary predictions due to an inherent lack of structural priors [4]. Thus, as shown in Fig. 1, we explore a cross-labeling strategy: labeling just three key slices per scan from the axial, sagittal and coronal views, respectively. This strategy capitalizes on the inter-slice similarity in 3D medical images while retaining the informative disparities across views [4] to facilitate 3D model training.

3 Experiments and Results

Materials. We perform extensive evaluation on the left atrium (LA) dataset [21] with 100 3D gadolinium-enhanced magnetic resonance images (GE-MRIs) and the brain tumor (BT) dataset [3] with 335 3D T2-FLAIR MRI. The images have the isotropic resolution of $0.625 \times 0.625 \times 0.625\ mm^3$ (LA) or $1 \times 1 \times 1$

Table 1. Comparison of the mean results over three runs. Standard deviations are in parentheses. $\star$ denotes cross-labeling. $*$ denotes $p \leq 0.05$ in pairwise comparison with our best version (†) via the Wilcoxon signed rank test. The best results are in **bold**.

Method	Setting			Metrics			
	Active Type	Labeled/Unlabeled	Labeled Slices	Dice (%) ↑	Jaccard (%) ↑	95HD (voxel) ↓	ASD (voxel) ↓
Left Atrium (LA) [21]							
Sup (cross; baseline)	random	20*/0	60	73.68 (10.81)*	59.36 (12.01)*	21.24 (4.90)*	6.14 (1.96)*
Sup (dense; upper bound)	random	80/0	7040	91.56 (2.06)	84.50 (3.49)	5.03 (1.62)	1.52 (0.44)
MT [15]	random	20*/60	60	76.45 (10.99)*	63.01 (12.65)*	19.75 (14.58)*	4.56 (1.93)*
UA-MT [26]	random	20*/60	60	80.09 (10.86)*	69.06 (12.45)*	13.47 (7.31)*	4.36 (3.06)*
CPS [6]	random	20*/60	60	66.84 (6.18)*	50.51 (6.83)*	18.26 (3.98)*	5.33 (1.26)*
ICT [17]	random	20*/60	60	78.69 (9.90)*	65.87 (12.27)*	13.85 (7.63)*	4.82 (3.04)*
CPCL [25]	random	20*/60	60	80.31 (6.03)*	67.51 (8.11)*	19.59 (6.69)*	5.73 (1.85)*
CAML [7]	random	20*/60	60	72.87 (11.70)*	58.49 (12.87)*	20.61 (12.44)*	5.57 (2.24)*
ACMT [24]	random	20*/60	60	77.24 (5.71)*	62.63 (7.13)*	29.47 (16.89)*	5.59 (1.35)*
DeSCO [4]	random	20*/60	60	79.25 (9.89)*	68.42 (10.52)*	21.24 (13.48)*	5.21 (2.06)*
FM-ABS (UA)	random	20*/60	60	83.68 (3.20)	72.07 (4.75)	21.62 (13.22)	5.14 (1.27)
FM-ABS (MV)	random	20*/60	60	84.93 (4.18)	74.04 (6.21)	16.92 (11.52)	3.89 (1.14)
FM-ABS (UA)	LC	20*/60	60	85.11 (3.84)	74.27 (5.76)	12.28 (5.08)	3.10 (1.44)
FM-ABS (MV)	LC	20*/60	60	85.86 (4.54)	75.49 (6.80)	**11.23** (4.77)	3.42 (1.26)
FM-ABS (UA)	HE	20*/60	60	85.12 (3.87)	74.28 (5.71)	14.22 (9.53)	3.54 (1.23)
FM-ABS (MV)	HE	20*/60	60	86.01 (3.04)	75.53 (3.04)	11.25 (5.54)	**2.88** (0.91)
FM-ABS (UA)	HER	20*/60	60	84.97 (7.19)	72.95 (9.40)	13.92 (4.73)	4.22 (1.56)
FM-ABS (MV)†	HER	20*/60	60	**86.14** (3.80)	**75.85** (5.79)	11.54 (5.14)	3.03 (0.75)
Brain Tumor (BT) [3]							
Sup (cross; baseline)	random	25*/0	75	67.21 (17.05)*	52.99 (18.65)*	14.12 (11.93)*	4.63 (2.83)*
Sup (dense; upper bound)	random	250/0	34173	87.07 (7.90)	77.48 (11.45)	7.84 (8.09)	1.79 (1.49)
MT [15]	random	25*/225	75	77.91 (17.92)*	66.86 (20.92)*	21.63 (24.71)*	2.50 (1.88)
UA-MT [26]	random	25*/225	75	78.43 (19.71)*	67.97 (21.58)*	19.07 (16.70)*	3.14 (3.57)*
CPS [6]	random	25*/225	75	77.98 (19.20)*	67.27 (21.63)*	18.81 (19.34)	2.75 (2.53)*
ICT [17]	random	25*/225	75	77.18 (19.26)*	66.22 (21.84)*	21.12 (25.38)*	2.70 (2.65)*
CPCL [25]	random	25*/225	75	78.95 (17.68)*	68.47 (20.48)*	17.14 (17.67)	3.02 (2.59)*
CAML [7]	random	25*/225	75	77.87 (14.65)*	65.92 (18.10)*	19.06 (21.32)*	2.38 (1.59)
ACMT [24]	random	25*/225	75	76.68 (19.83)*	65.70 (22.16)*	19.81 (18.14)*	3.23 (2.96)*
DeSCO [4]	random	25*/225	75	75.32 (16.35)*	64.08 (22.37)*	22.34 (19.22)*	3.19 (3.05)*
FM-ABS (UA)	random	25*/225	75	80.02 (17.83)	69.62 (20.56)	20.79 (25.59)	2.26 (1.99)
FM-ABS (MV)	random	25*/225	75	80.12 (15.23)	70.05 (18.79)	19.81 (23.42)	2.31 (1.16)
FM-ABS (UA)	LC	25*/225	75	80.89 (15.21)	70.24 (18.44)	17.99 (21.41)	2.10 (1.43)
FM-ABS (MV)	LC	25*/225	75	80.99 (14.93)	70.63 (17.86)	16.87 (20.69)	**1.98** (1.57)
FM-ABS (UA)	HE	25*/225	75	80.62 (15.73)	70.01 (19.03)	**13.80** (17.17)	2.42 (2.09)
FM-ABS (MV)	HE	25*/225	75	80.77 (16.86)	70.36 (19.91)	16.58 (20.43)	2.52 (2.73)
FM-ABS (UA)	HER	25*/225	75	81.78 (15.90)	71.26 (19.01)	16.05 (19.92)	2.27 (1.70)
FM-ABS (MV)†	HER	25*/225	75	**81.93** (13.53)	**71.78** (17.13)	17.63 (21.26)	2.26 (1.58)

mm^3 (BT). We follow the data split and preprocessing used in [26, 25]. For LA, 80 and 20 samples are for training and testing, respectively [26]. For BT, 250, 25 and 60 samples are used for training, validation and testing, respectively [25].

Implementation and Evaluation Metrics. The framework is implemented on PyTorch using an NVIDIA A100 GPU. Following the previous semi-supervised works [26, 24], the same 3D V-Net [14] is adopted as the specialist. Considering efficiency, we adopt the open-source MobileSAM [27] as the generalist ($12ms$ per slice) with lightweight image encoder ViT-Tiny [19]. Other SAM variants are also discussed later. We randomly crop patches of $112 \times 112 \times 80$ (LA) or $96 \times 96 \times 96$ (BT) voxels as the input and use sliding window strategy with stride of $18 \times 18 \times 4$ (LA) or $64 \times 64 \times 64$ (BT) voxels for inference. The batch size is set to 4 including 2 cross-labeled data and 2 unlabeled data. $t_{\max}$ is set to 20,000. The learning rate is initialized as 0.01 and decayed with a power of 0.9 after each iteration. Weak data augmentation, including randomly flipping and rotating, is applied [26]. The Dice score (Dice), Jaccard, 95% Hausdorff Distance (95HD) and Average Surface Distance (ASD) are adopted as our evaluation metrics. Code will be available at `https://github.com/lemoshu/FM-ABS`.

Comparison Study. Table 1 presents the quantitative results on the two datasets under extremely limited labeling budgets. Besides the supervised baselines (Sup), we include recent top-performing SSL methods [15, 26, 6, 17, 25, 7, 24, 4]. All methods are implemented with the same backbone and training protocols to ensure

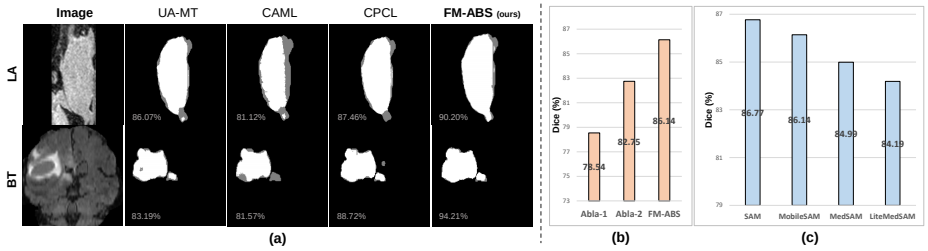

Fig. 3. (a) Exemplar 2D results. Grey color indicates the mismatch between the prediction and the ground truth. (b) Ablation studies. (c) Sensitivity on different generalists.

fairness. As observed, under the standard random sampling protocol, both variants of FM-ABS consistently outperform the supervised baselines and recent SSL methods, demonstrating the effectiveness of our specialist-generalist collaboration in the barely supervised context. Generally, employing majority voting (MV) for calibrating multi-view generalist-based pseudo labels yields superior outcomes. The performance of FM-ABS is further enhanced by adopting model-informed active selection, highlighting the benefits of expert-model collaboration. Specifically, in the LA task, FM-ABS (MV) equipped with HER sampling strategy attains a Dice score of 86.14%, trailing the supervised upper bound by 5.42% Dice but with merely 0.8% of slices labeled. For the BT task, FM-ABS (MV) also prefers the HER sampling strategy, achieving the best performance with an 81.93% Dice score, utilizing merely 0.02% of labeled slices relative to the supervised upper bound (87.07% Dice). Fig. 3 (a) further shows that the predictions of our method align more accurately with the ground truth.

Ablation Study and Sensitivity against Different Generalists. To evaluate the effectiveness of each component, we perform an ablation study using our best version in the LA task, as depicted in Fig. 3 (b). Firstly, eliminating $\mathcal{L}_{sup}^{SAM}$ (Abla-1) results in a 7.6% decrease in Dice score, highlighting the significant contribution of the specialist-collaborated generalist-based pseudo supervision. In Abla-2, the removal of $\mathcal{L}_c$ reveals the crucial role of the classical consistency constraint under perturbations. Such stability learning encourages isotropic local smoothing around each data point [16], enhancing the model's generalizability and forcing it to learn from the intrinsic data structure that could alleviate confirmation bias induced by noisy generalist-based labels. Beyond MobileSAM [27] with an image encoder of 5.78M parameters, we also explore other SAM variants: Meta's SAM (ViT-B) with 86M parameters, the medically-specialized MedSAM (86M) [13], and its lightweight counterpart, LiteMedSAM (5.78M). Fig. 3 (c) reveals that SAM (ViT-B) marginally outperforms MobileSAM, albeit at the cost of slower inference for pseudo labeling. Interestingly, MedSAM, despite being specifically fine-tuned on extensive medical images, yields degraded performance, possibly due to over-specialization on the finetuning dataset.

4 Conclusion

In this study, we proposed the Foundation Model-driven Active Barely Supervised (FM-ABS) learning paradigm for developing specialist 3D medical image segmentation models with meager annotation budgets. FM-ABS modernizes previous semi-supervised learning with two key data-centric designs: (i) leveraging the class-specific prompts derived from the in-training specialist model to prompt the frozen generalist model with strong zero-shot generalizability for complementary noisy-yet-effective pseudo supervision, and (ii) active cross-labeling with notably low annotation efforts to progressively provide the specialist model with informative and efficient supervision to meet model-informed needs, which benefits the automatic prompt generation in turn. Our experiments demonstrated the superiority of our approach over previous state-of-the-art methods. Notably, grounded in the data-centric spirit, our approach is expected to be flexibly extended to other backbones and SSL frameworks beyond those used in our study.

Acknowledgments. This research was partly supported by General Research Fund (No. 14205419) and Hong Kong PhD Fellowship from Research Grants Council of Hong Kong.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac MR image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 253–260. Springer (2017)
2. Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., Liu, T.: Understanding and improving early stopping for learning with noisy labels. Advances in Neural Information Processing Systems **34**, 24392–24403 (2021)
3. Bakas, S.: BraTS MICCAI brain tumor dataset (2020). https://doi.org/10.21227/hdtd-5j88
4. Cai, H., Li, S., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Orthogonal annotation benefits barely-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3302–3311 (2023)
5. Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. arXiv preprint arXiv:2309.08842 (2023)
6. Chen, X., Yuan, Y., Zeng, G., Wang, J.: 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)
7. Gao, S., Zhang, Z., Ma, J., Li, Z., Zhang, S.: Correlation-aware mutual learning for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 98–108. Springer (2023)

8. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
11. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Transactions on Medical Imaging **37**(12), 2663–2674 (2018)
12. Luo, X., Chen, J., Song, T., Chen, Y., Wang, G., Zhang, S.: Semi-supervised medical image segmentation through dual-task consistency. In: AAAI Conference on Artificial Intelligence (2021)
13. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision. pp. 565–571. IEEE (2016)
15. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. pp. 1195–1204 (2017)
16. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning **109**(2), 373–440 (2020)
17. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Networks **145**, 90–106 (2022)
18. Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N., Zhang, S.: A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. IEEE Transactions on Medical Imaging **39**(8), 2653–2663 (2020)
19. Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: TinyViT: Fast pretraining distillation for small vision Transformers. In: European Conference on Computer Vision. pp. 68–85. Springer (2022)
20. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 297–306. Springer (2021)
21. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical Image Analysis **67**, 101832 (2021)
22. Xu, Z., Lu, D., Luo, J., Zheng, Y., Tong, R.K.y.: Separated collaborative learning for semi-supervised prostate segmentation with multi-site heterogeneous unlabeled mri data. Medical Image Analysis **93**, 103095 (2024)
23. Xu, Z., Lu, D., Yan, J., Sun, J., Luo, J., Wei, D., Frisken, S., Li, Q., Zheng, Y., Tong, R.K.y.: Category-level regularized unlabeled-to-labeled learning for semi-supervised prostate segmentation with multi-site unlabeled data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–13. Springer (2023)

24. Xu, Z., Wang, Y., Lu, D., Luo, X., Yan, J., Zheng, Y., Tong, R.K.y.: Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. Medical Image Analysis p. 102880 (2023)
25. Xu, Z., Wang, Y., Lu, D., Yu, L., Yan, J., Luo, J., Ma, K., Zheng, Y., Tong, R.K.y.: All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(7), 3174–3184 (2022)
26. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 605–613. Springer (2019)
27. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. arXiv preprint arXiv:2306.14289 (2023)
28. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer (2017)
29. Zhao, Z., Xu, K., Li, S., Zeng, Z., Guan, C.: MT-UDA: Towards unsupervised cross-modality medical image segmentation with limited source labels. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 293–303. Springer (2021)