Source-Free Domain Adaptation for Medical Image Segmentation via LLM-Agent Collaboration

Tianyu Zhang[1⋆], Jianhang Ji[2⋆], Zhiming Cheng[1], Jianxiang Zhao[1], Fengjun Xiao[1], and Shuai Wang[1(✉)]

[1] Hangzhou Dianzi University, Hangzhou, China
[2] City University of Macau, Macao, China
shuaiwang.tai@gmail.com

Abstract. Source-Free Domain Adaptation (SFDA) is imperative for privacy-preserving medical image analysis, allowing models to adapt to target domains without accessing source data. However, existing SFDA methods face fundamental limitations: entropy-based approaches (e.g., TENT, SAR) only update limited model parameters and suffer from instability; self-training or contrastive consistency methods (e.g., CCRC) still rely on the source model for pseudo-label generation, inevitably accumulating noise under domain shift; and even recent SAM-based correction (e.g., IPLC) depends on a single foundation model without semantic-level verification. In this paper, we propose **RaMA**, a Reliability-Aware Multi-Agent Adaptation Framework that introduces Large Language Models (LLMs) into the SFDA loop. Specifically, we leverage a Multimodal LLM as a Cognitive Meta-Controller to perform semantic error checking and generate spatial constraints, which guide a Multi-Agent System (MAS) of diverse SAMs to produce consensus-based pseudo-labels. Furthermore, an LLM-derived reliability score dynamically weights the self-training loss, suppressing noisy corrections from propagating. Extensive experiments on the M&Ms cardiac segmentation benchmark show that RaMA outperforms all compared SFDA methods by a large margin, improving average Dice by 5.58% and reducing ASSD by 50% over the best existing approach.

Keywords: Source-Free Domain Adaptation · Large Language Models · Medical Image Segmentation · Multi-Agent System.

1 Introduction

Medical image segmentation plays a pivotal role in clinical practice, enabling quantitative analysis of anatomical structures for disease diagnosis, surgical planning, and treatment monitoring [22]. With the advent of deep learning, automated segmentation methods [6,10,11,23] have achieved remarkable performance, substantially reducing the burden of manual annotation. However, these

⋆ Equal contribution as first authors

models are typically trained on data from a single institution or scanner (i.e., the source domain) and suffer significant performance degradation when deployed to other clinical sites with different scanners or imaging protocols (i.e., the target domain), a problem known as domain shift. Unsupervised Domain Adaptation (UDA) [8,28] mitigates this by aligning feature distributions between the source and target domains, but it requires simultaneous access to both source domain data and target domain data during training. In clinical practice, stringent privacy regulations (e.g., HIPAA, GDPR) often prohibit sharing of source patient data across institutions [9], and the sheer volume of medical imaging data also makes cross-site transmission and storage prohibitively costly. These constraints necessitate Source-Free Domain Adaptation (SFDA) [2,15,30], which adapts the model using only a pre-trained source model and unlabeled target data.

To address the above challenges, various SFDA methods have been proposed. Bateson et al. [2] introduced a source-relaxed constraint framework using entropy loss guided by anatomical priors for medical image segmentation. SHOT [15] proposed a classic paradigm combining information maximization with self-supervised pseudo-labeling to align target features to source prototypes. TENT [25] proposed an online adaptation strategy by updating Batch Normalization parameters via entropy minimization, with subsequent improvements by EATA [20], SAR [21], and STTA [14] for better efficiency, stability, and cross-modality medical adaptation. More recently, Ma et al. [18] designed a contrastive class-relationship consistency (CCRC) framework that leverages teacher-student class relationship matrices for cross-modality cardiac SFDA. However, these methods rely solely on the source model to generate pseudo-labels, which are inevitably noisy under domain shift, leading to confirmation bias and error accumulation [16,29,32]. To introduce external correction, IPLC [26] pioneered the use of SAM [12,27] to iteratively refine pseudo-labels, marking the first attempt to incorporate a foundation model into the SFDA loop. Despite this advance, IPLC still has notable limitations: (1) it depends on a single SAM variant, which is susceptible to model-specific hallucinations without cross-model validation; (2) it lacks high-level semantic reasoning to detect anatomically implausible predictions (e.g., fragmented organs or impossible spatial relationships); and (3) it treats all refined pseudo-labels equally without assessing their reliability, allowing noisy corrections to propagate through self-training. Meanwhile, recent advances in reasoning segmentation [13], LLM-boosted segmentation [24], and collaborative multi-agent medical analysis [17] suggest that LLMs can serve as intelligent controllers for complex visual tasks, yet their integration into the SFDA pipeline remains unexplored.

To address these challenges, we propose **RaMA**, a Reliability-Aware Multi-Agent framework that shifts the paradigm of pseudo-label refinement from signal denoising to cognitive reasoning. Specifically, we employ a Multimodal LLM as a "Meta-Controller" to inspect segmentation failures and generate Spatial Constraints (points/boxes). These constraints guide a collaborative team of heterogeneous agents (SAM3, MedSAM2 [19], SAM-Med2D [7]) to produce consensus-based corrections, mitigating single-model hallucinations. To prevent learning

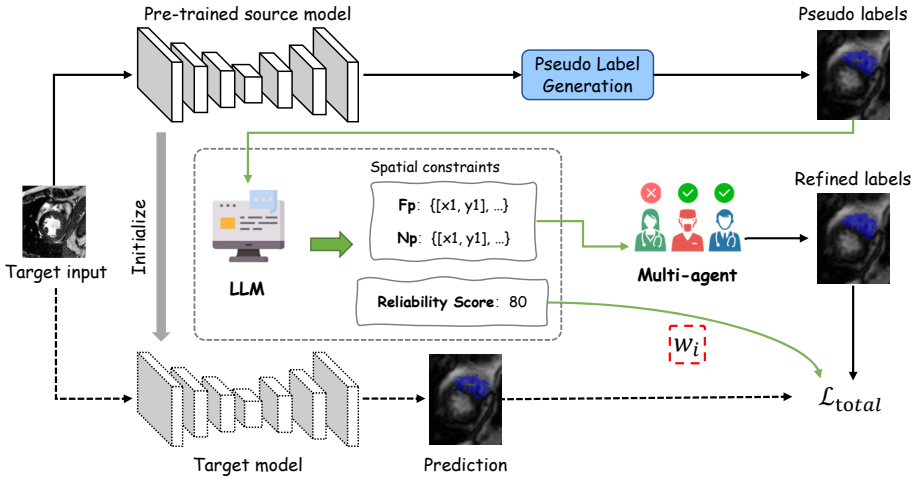

Fig. 1. Overview of the RaMA Framework. An LLM acts as a cognitive controller to generate spatial constraints for diverse SAM agents. FP (False Positive) denotes background regions misclassified as foreground; FN (False Negative) denotes foreground regions missed by the model. The resulting consensus mask and reliability score guide the iterative adaptation of the target model.

from unreliable samples [5], we further introduce a reliability-aware loss weighted by the LLM's confidence and agent consensus. Our contributions are three-fold: (1) We propose an LLM-driven refinement mechanism that translates semantic reasoning into geometric constraints for automated segmentation correction. (2) We design a Heterogeneous Multi-Agent System with a reliability-aware weighting strategy to robustly suppress noisy pseudo-labels during iterative self-training. (3) Extensive experiments on the M&Ms benchmark demonstrate that RaMA achieves state-of-the-art SFDA performance, outperforming all compared methods across multiple target domains.

2 Method

2.1 Problem Formulation

We address the challenging problem of Source-Free Domain Adaptation (SFDA) for medical image segmentation. Let $\mathcal{D}_S = \{(x_s^i, y_s^i)\}_{i=1}^{N_S}$ denote the labeled source domain, and $\mathcal{D}_T = \{x_t^j\}_{j=1}^{N_T}$ denote the unlabeled target domain, where $x \in \mathbb{R}^{H\times W}$ represents the input image and $y \in \{0,1\}^{H\times W}$ is the binary segmentation mask. The source and target domains share the same label space but differ in data distributions, i.e., $P(\mathcal{D}_S) \neq P(\mathcal{D}_T)$, due to variations in imaging protocols or scanner vendors.

We assume a source model f_{θ_S}, parameterized by θ_S, has been pre-trained on $\mathcal{D}_S$ and is provided to the target domain. In the SFDA setting, access to the

source data $\mathcal{D}_S$ is strictly prohibited to preserve patient privacy. Our objective is to learn a target-adapted model f_{θ_T} initialized with θ_S, utilizing only the unlabeled target data $\mathcal{D}_T$, such that it minimizes the segmentation error on $\mathcal{D}_T$. The overall framework, RaMA (Reliability-Aware Multi-Agent Adaptation), is illustrated in Fig. 1. It consists of three synergistic modules: (1) LLM-driven spatial constraint generation, (2) Multi-Agent collaborative refinement, and (3) Reliability-aware self-training.

2.2 LLM-Driven Spatial Constraint Generation

Standard pseudo-label refinement methods often rely on pixel-level uncertainty or entropy, which struggle to correct semantic-level errors (e.g., missing anatomical components or anatomically impossible shapes). We propose to inject high-level semantic knowledge via a Multimodal Large Language Model (LLM), specifically leveraging its reasoning capabilities to "diagnose" segmentation failures.

Given a target image $x_t \in \mathcal{D}_T$, we first generate an initial pseudo-label $\hat{y}_{init} = f_{\theta_T}(x_t)$ using the current model. To facilitate LLM understanding, we generate a visual overlay $I_{vis} = \text{Overlay}(x_t, \hat{y}_{init})$ where the segmentation mask is superimposed on the original image with high contrast. The LLM functions as a "Cognitive Meta-Controller" prompted to act as a senior radiologist. We design a hierarchical structured prompt $\mathcal{P}_{text}$ covering three dimensions: (1) a role definition that establishes the LLM as an expert radiologist specializing in cardiac segmentation; (2) a task description instructing it to identify False Negatives (FN) where the model missed the organ and False Positives (FP) where background is mistaken for the organ; and (3) an actionable output format requiring a JSON list of corrective point coordinates for under-segmented and over-segmented areas.

The LLM processes the visual-language input and outputs a set of spatial constraints $\mathcal{C}$ (using Qwen-VL [1]):

$$\mathcal{C} = \text{LLM}(I_{vis}, \mathcal{P}_{text}) = \{(p_k, l_k)\}_{k=1}^{K} \tag{1}$$

where $p_k \in \mathbb{R}^2$ represents the spatial coordinate of the k-th prompt point, and $l_k \in \{0, 1\}$ serves as the class indicator (1 for positive/foreground click, 0 for negative/background click). This process effectively translates abstract semantic reasoning into precise geometric constraints compatible with promptable segmentation models.

2.3 Collaborative Multi-Agent Refinement

Relying on a single foundation model for refinement poses a risk of "hallucination," where the model might generate plausible but incorrect structures. To mitigate this, we orchestrate a Multi-Agent System (MAS) that aggregates insights from diverse experts with complementary inductive biases. Specifically, we employ three distinct agents: $\mathcal{A}_{sam3}$ (SAM 3 [4]), a general-purpose vision model with superior boundary delineation capabilities pre-trained on massive natural

image datasets; $\mathcal{A}_{med}$ (MedSAM 2 [19,31]), a medical-specific model fine-tuned on diverse medical modalities, offering strong priors for biological shapes and organ connectivity; and $\mathcal{A}_{slice}$ (SAM-Med2D [7]), a specialized model optimized for slice-wise consistency in volumetric data.

Each agent $\mathcal{A}_m$ independently predicts a refined candidate mask $\hat{y}_m$ guided by the shared spatial constraints $\mathcal{C}$:

$$\hat{y}_m = \mathcal{A}_m(x_t, \mathcal{C}), \quad m \in \{sam3, med, slice\} \tag{2}$$

To fuse these predictions, we employ a pixel-wise majority voting mechanism. This filters out stochastic errors and model-specific outliers, ensuring the final pseudo-label represents a high-confidence consensus:

$$\hat{y}_{ref}^{(u,v)} = \mathbb{I}\left(\frac{1}{M}\sum_{m}\hat{y}_m^{(u,v)} > \tau_{vote}\right) \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function, (u, v) denotes pixel coordinates, $M = 3$ is the number of agents, and $\tau_{vote} = 0.5$ is the voting threshold.

2.4 Reliability-Aware Self-Training

Blindly retraining on all refined pseudo-labels can lead to noise accumulation, as even the multi-agent system may fail on hard samples with severe artifacts. To address this confirmation bias and potential LLM hallucinations [5], we introduce a reliability-aware training strategy to select pseudo-labels and modulate the contribution of training samples.

Based on the LLM quality scores, we adopt a tri-partition strategy for pseudo-label management. Concretely, for each slice i, let $S_{i,c}$ denote the per-organ score. Given a low threshold τ_l and a high threshold τ_h, the slice is discarded if $\exists c,\ S_{i,c} < \tau_l$; retained without modification if $\forall c,\ S_{i,c} \geq \tau_h$; and otherwise forwarded to the multi-agent refinement module for consensus-based correction.

For the retained or refined pseudo-label $\hat{y}_{ref}^i$, we define a sample-wise reliability weight w_i by averaging the normalized organ-level scores:

$$w_i = \frac{1}{C}\sum_{c=1}^{C}\frac{S_{i,c}}{100}, \tag{4}$$

where $C = 3$ denotes the number of target organs. A high w_i indicates that the pseudo-label is more reliable according to the LLM assessment.

Finally, we update the target model parameters θ_T by minimizing a reliability-weighted objective function:

$$\mathcal{L}_{total} = \frac{1}{B}\sum_{i=1}^{B} w_i \cdot \mathcal{L}_{Dice}(\hat{y}_{ref}^i, f_{\theta_T}(x_i)) + \beta\mathcal{L}_{curv}, \tag{5}$$

where $\mathcal{L}_{Dice}$ is the Dice loss and $\mathcal{L}_{curv}$ is a curvature regularizer that penalizes irregular boundaries. By weighting each sample by its LLM-assessed reliability

w_i, the model prioritizes trustworthy pseudo-labels and suppresses the propagation of noisy corrections. After this initial reliability-weighted adaptation, subsequent rounds employ self-training with confidence-based pseudo-label filtering to further refine the model.

3 Experiments and Results

3.1 Experimental Details

Datasets and Preprocessing. We evaluated our method on the M&Ms (Multi-Centre, Multi-Vendor, and Multi-Disease) cardiac segmentation dataset [3], which was acquired from four scanner vendors: Siemens (Vendor A), Philips (Vendor B), General Electric (Vendor C), and Canon (Vendor D), containing 192, 252, 150, and 100 cardiac MRI volumes, respectively. Following standard SFDA protocols [18, 26], Vendor A was used as the labeled source domain, while Vendors B, C, and D served as unlabeled target domains. The segmentation targets are the Left Ventricle (LV), Right Ventricle (RV), and Myocardium (MYO). All 2D short-axis slices were normalized to $[-1, 1]$ and center-cropped to 256×256 pixels. In each target domain, data was randomly split into 70% for training (labels discarded), 10% for validation, and 20% for testing.

Implementation Details. We implemented the framework using PyTorch on a single NVIDIA GeForce RTX 3090 GPU. The segmentation backbone is a ResUnet with a ResNet-34 encoder (pretrained on ImageNet), with approximately 24M parameters.

We employed Qwen-VL [1] to assess the quality of initial pseudo-labels. The LLM assigns a quality score (0–100) to each organ and generates point prompts. We applied a tri-partition filtering strategy based on the score threshold τ: slices where any organ scored < 40 were discarded to prevent noise accumulation; samples scoring between 40 and 90 were refined using our Multi-SAM ensemble, where SAM3, MedSAM2, and SAM-Med2D perform pixel-level voting guided by LLM-generated positive/negative points (a fallback mechanism rejects refinements whose Dice similarity with the original pseudo-label falls below 0.5); and high-confidence samples where all organ scores are at least 90 were used directly without modification.

The adaptation process is divided into multiple rounds. In Round 0, the source model is adapted on the target domain using the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay $1e^{-4}$) with a linear warmup and Cosine Annealing scheduler (minimum LR = $0.01\times$ initial LR). The objective combines an LLM-score-weighted Dice loss with the curvature regularizer ($\beta = 0.01$), as described in Sec. 2.4. In Rounds 1–3, pseudo-labels were regenerated by the previous best model. Slices with mean pixel confidence below 0.6 were filtered out, and DBSCAN was applied to retain the principal cardiac ROI. The model was fine-tuned with a reduced learning rate (3×10^{-5}) and without LLM-score weighting. To enhance robustness, we applied extensive data augmentations, including random flip, rotation (90°/180°/270°), scaling (0.8–1.2), and intensity jittering.

Table 1. Quantitative comparison on the M&Ms target domains. Per-structure Dice (%) and ASSD (mm) are reported as test-slice mean±standard deviation; Avg denotes the mean across structures. Bold values indicate the best performance.

Target	Method	Dice (%) ↑				ASSD (mm) ↓			
		LV	MYO	RV	Avg	LV	MYO	RV	Avg
B	Source Only	86.30±18.33	75.91±14.70	77.65±22.59	79.95	4.06±11.34	4.25±10.91	4.88±7.87	4.40
	TENT [25]	87.59±15.65	76.95±11.84	78.48±21.49	81.01	4.83±17.24	4.00±14.87	5.05±11.17	4.63
	EATA [20]	85.75±21.46	74.85±17.53	77.47±24.52	79.36	8.23±31.04	7.10±27.89	9.42±26.01	8.25
	SAR [21]	87.54±15.62	77.01±11.62	78.33±21.54	80.96	4.65±17.00	3.76±14.49	4.97±10.88	4.46
	CCRC [18]	87.69±15.23	77.60±12.20	79.29±20.60	81.53	2.68±7.51	2.16±6.81	3.86±**5.69**	2.90
	IPLC [26]	86.85±13.50	76.52±13.10	77.81±23.00	80.39	2.69±9.10	2.05±7.80	3.04±9.70	2.59
	Ours	**90.49±10.73**	**82.94±5.67**	**83.61±17.70**	**85.68**	**1.01±0.75**	**1.04±0.45**	**2.73**±11.15	**1.59**
C	Source Only	80.13±21.93	63.95±20.00	65.68±28.57	69.92	6.20±15.37	7.57±16.29	8.14±10.92	7.30
	TENT [25]	81.24±20.18	64.88±18.87	67.44±27.65	71.19	11.76±25.85	11.18±24.78	9.18±16.35	10.71
	EATA [20]	78.99±22.01	61.19±20.86	66.79±27.31	68.99	17.89±30.03	19.31±29.56	13.65±21.18	16.95
	SAR [21]	80.34±21.65	63.87±19.76	66.71±28.03	70.31	12.89±27.18	12.98±26.67	9.47±17.00	11.78
	CCRC [18]	82.72±15.31	68.18±16.17	67.78±28.86	72.89	5.70±10.40	5.23±10.03	5.43±**9.95**	5.45
	IPLC [26]	83.19±19.10	66.96±18.90	66.91±28.40	72.35	6.60±20.50	6.13±17.60	6.46±12.20	6.40
	Ours	**90.53±10.55**	**76.42±12.35**	**79.45±21.86**	**82.13**	**1.92±8.76**	**2.04±8.74**	**4.05**±14.91	**2.67**
D	Source Only	83.40±17.81	68.96±16.46	73.81±24.14	75.39	7.90±18.22	6.71±14.48	10.94±21.12	8.52
	TENT [25]	84.04±15.86	69.43±15.50	75.74±21.64	76.40	9.95±21.04	7.39±16.78	10.62±20.08	9.32
	EATA [20]	81.80±20.08	66.85±19.05	73.87±24.35	74.17	15.61±29.24	13.40±26.06	18.20±29.62	15.74
	SAR [21]	84.12±15.58	68.90±15.42	75.69±**21.36**	76.24	9.53±20.61	7.29±16.94	9.77±18.72	8.86
	CCRC [18]	83.88±16.05	70.92±14.45	76.88±22.41	77.23	8.25±17.92	5.91±11.25	8.90±**17.68**	7.69
	IPLC [26]	84.74±15.80	71.79±14.60	75.27±23.90	77.27	7.41±17.70	4.55±10.30	9.25±20.10	7.07
	Ours	**88.12±11.44**	**75.58±11.89**	**78.04**±24.71	**80.58**	**2.01±8.48**	**1.97±8.47**	**7.15**±23.33	**3.71**

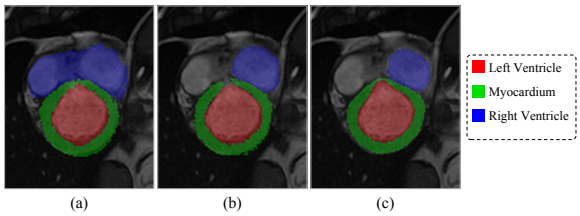

Fig. 2. Qualitative pseudo-label refinement on Vendor C. From left to right: (a) initial pseudo-label, (b) Multi-SAM consensus pseudo-label, and (c) ground truth. Red, green, and blue denote LV, MYO, and RV, respectively.

3.2 Experimental Results

Comparison with State-of-the-Art. We compare our method with five SFDA methods: TENT [25], EATA [20], SAR [21], CCRC [18], and IPLC [26]. We adopt the Dice score and Average Symmetric Surface Distance (ASSD) for evaluation. Table 1 reports per-structure results for each target domain. Our method achieves consistent improvements across all target domains and cardiac structures. On Vendor B, the overall Dice improves from 81.53% (CCRC) to 85.68%, with the largest per-structure gain on MYO (+5.34%) where thin-walled anatomy demands precise boundary delineation. On Vendor C, our framework achieves 82.13% average Dice (+9.24% over CCRC), with RV Dice improving by 11.67% over the next best method, demonstrating the effectiveness of external semantic correction when data is scarce. On Vendor D, our method reduces av-

Table 2. Component ablation with a direct-inference baseline. "Direct" applies LLM-guided Multi-SAM consensus without target-model fine-tuning or self-training. Avg Dice (%) and Avg ASSD (mm) across LV, MYO, and RV are reported.

Setting	Vendor B		Vendor C		Vendor D		Average	
	Dice↑	ASSD↓	Dice↑	ASSD↓	Dice↑	ASSD↓	Dice↑	ASSD↓
Direct Inference on Target Test Set								
LLM + Multi-SAM	77.88	2.54	68.36	3.87	75.76	3.37	74.00	3.26
Component Ablation								
Source Only	79.95	4.40	69.92	7.30	75.39	8.52	75.09	6.74
+ Multi-Agent Refinement	84.09	1.63	78.87	3.53	78.43	5.41	80.46	3.52
+ Reliability Weighting	**85.68**	**1.59**	**82.13**	**2.67**	**80.58**	**3.71**	**82.80**	**2.66**

erage ASSD from 7.07mm (IPLC) to 3.71mm. Entropy-based methods (TENT, SAR) provide marginal Dice gains over the source model but often increase ASSD, while EATA suffers from catastrophic forgetting. Methods with structured adaptation (CCRC, IPLC) achieve better ASSD, yet remain limited by source-model-generated pseudo-labels. Our LLM-guided multi-agent refinement breaks this ceiling by introducing external semantic reasoning and cross-model consensus.

Ablation Study. To investigate the contribution of each component, we conducted ablation experiments across all three target domains. Table 2 reports average Dice and ASSD when progressively adding modules. Direct inference with LLM-guided Multi-SAM consensus achieves 74.00% Dice and 3.26mm ASSD, whereas the full framework reaches 82.80% Dice and 2.66mm ASSD, showing that consensus refinement is most effective when integrated with reliability-aware target-domain adaptation rather than used as standalone post-processing. Figure 2 qualitatively illustrates the pseudo-label improvement achieved by multi-agent refinement. Starting from the source model without adaptation (Source Only, 75.09% Dice, 6.74mm ASSD), performance is severely limited by domain shift. Introducing Multi-Agent Refinement brings +5.37% Dice and reduces ASSD by 47.78% (6.74→3.52mm). This validates the core design of Sec. 2.2–2.3: the LLM identifies semantic-level failures (e.g., missed RV regions, fragmented MYO) that the source model cannot self-correct, while the heterogeneous multi-agent consensus filters out individual model hallucinations, with each agent contributing complementary strengths in boundary delineation, medical shape priors, and slice-wise consistency. Adding reliability-aware training improves Dice from 80.46% to 82.80% and reduces ASSD from 3.52mm to 2.66mm. The gain is most pronounced on Vendor C (+3.26% Dice), where the initially severe domain gap benefits most from progressive refinement: as the adapted model improves across rounds, its pseudo-labels become more accurate, creating a virtuous cycle that gradually recovers fine-grained boundary details.

4 Conclusion

We presented RaMA, a novel SFDA framework that integrates LLM-driven semantic reasoning with a heterogeneous multi-agent ensemble for reliable pseudo-label refinement. By combining the cognitive capabilities of LLMs with the complementary segmentation strengths of diverse foundation models, our method effectively mitigates the noisy pseudo-label bottleneck in source-free settings. Extensive experiments on the M&Ms benchmark demonstrate consistent improvements over existing methods across multiple target domains, with average Dice gains of 5.58% and ASSD reductions of over 50% compared to the best prior approach. In future work, we plan to extend the framework to more imaging modalities and clinical scenarios to further validate its generalizability.

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

Acknowledgements. This research was supported by National Natural Science Foundation of China under Grant No. 62571173 and Zhejiang Provincial Natural Science Foundation of China under Grant No. LD25F020005.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I.: Source-relaxed domain adaptation for image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 490–499. Springer (2020)
3. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021)
4. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
5. Chen, J., Yang, D., Wu, T., Jiang, Y., Hou, X., Li, M., Wang, S., Xiao, D., Li, K., Zhang, L.: Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
8. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE transactions on medical imaging **39**(7), 2415–2425 (2020)

9. Guan, H., Yap, P.T., Bozoki, A., Liu, M.: Federated learning for medical image analysis: A survey. Pattern recognition **151**, 110424 (2024)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
11. 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)
12. 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. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
13. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9579–9589 (2024)
14. Li, X., Fang, H., Wang, C., Liu, M., Duan, L., Xu, Y.: Cache-driven spatial test-time adaptation for cross-modality medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 146–156. Springer (2024)
15. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International conference on machine learning. pp. 6028–6039. PMLR (2020)
16. Liu, Y., Ye, W., Liu, H., Chen, Z., Li, P., Xu, R.X., Sun, M.: Harp: Harmonization and adaptive refinement of pseudo-labels for cross-domain medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 311–321. Springer (2025)
17. Lyu, X., Liang, Y., Chen, W., Ding, M., Yang, J., Huang, G., Zhang, D., He, X., Shen, L.: Wsi-agents: A collaborative multi-agent system for multi-modal whole slide image analysis. arXiv preprint arXiv:2507.14680 (2025)
18. Ma, A., Zhu, Q., Li, J., Nielsen, M., Chen, X.: Source-free domain adaptation for cross-modality cardiac image segmentation with contrastive class relationship consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 574–583. Springer (2025)
19. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
20. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022)
21. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. arXiv preprint arXiv:2302.12400 (2023)
22. Rayed, M.E., Islam, S.S., Niha, S.I., Jim, J.R., Kabir, M.M., Mridha, M.F.: Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in medicine unlocked **47**, 101504 (2024)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
24. Tang, F., Ma, W., He, Z., Tao, X., Jiang, Z., Zhou, S.K.: Pre-trained llm is a semantic-aware and generalizable segmentation booster. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 402–412. Springer (2025)

25. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
26. Zhang, G., Qi, X., Yan, B., Wang, G.: Iplc: iterative pseudo label correction guided by sam for source-free domain adaptation in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 351–360. Springer (2024)
27. Zhang, Y., Jiao, R.: Towards segment anything model (sam) for medical image segmentation: a survey. arXiv preprint arXiv:2305.03678 (2023)
28. Zhou, W., Ji, J., Cui, W., Wang, Y., Yi, Y.: Unsupervised domain adaptation fundus image segmentation via multi-scale adaptive adversarial learning. IEEE journal of biomedical and health informatics **28**(10), 5792–5803 (2023)
29. Zhou, W., Ji, J., Cui, W., Yi, Y.: Pseudo-label clustering-driven dual-level contrast learning based source-free domain adaptation for fundus image segmentation. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 492–503. Springer (2023)
30. Zhou, W., Ji, J., Cui, W., Yi, Y., Chen, Y.: Diffusion-driven dual-flow source-free domain adaptation for medical image segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 4082–4087. IEEE (2024)
31. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)
32. Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5982–5991 (2019)