BiLD: Generalizable Bilevel Knowledge Distillation from Vision Foundation Models for 3D Medical Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Generalizable 3D medical image segmentation is increasingly vital for clinical deployment, given cross-site distribution shifts and restricted access to target-site data. However, cross-site distribution shifts and restricted access to target-site data during training make generalization to unseen sites challenging. Vision foundation models (VFMs) offer transferable segmentation priors, but their computational demands limit direct deployment. Existing distillation approaches transfer these priors to lightweight students through teacher adaptation (TA) and student distillation (SD). However, optimizing these tasks primarily for source-domain fitting does not explicitly align knowledge transfer with the student's cross-site generalization. We propose Generalizable Bilevel Knowledge Distillation (BiLD), a framework that couples lower-level TA with upper-level SD. A dynamic response strategy incorporates the response of teacher adaptation into student updates using a Gauss–Newton approximation and conjugate-gradient iterations. To support this coupled optimization without target-site data, we introduce a pseudo-site generator that constructs diverse appearance variations from source-domain volumes through acquisition-motivated perturbations while preserving anatomical labels. These pseudo-sites expose the upper-level distillation objective to simulated distribution shifts, encouraging knowledge transfer beyond source-domain matching.Experiments on brain and cardiac segmentation across five external datasets demonstrate improved zero-shot cross-site generalization, with mean out-of-domain Dice exceeding the strongest baseline in each task by 3.23% and 5.99%, respectively. Two target-domain fine-tuning experiments further utperforms the strongest baseline in each task by 2.25% and 3.37%in Dice for brain and cardiac, respectively. Code will be released upon acceptance.

1 Introduction

Precise medical image segmentation (MIS) plays a crucial role in many medical analyses, such as tumor detection Elmahdy et al. (2021); Soomro et al. (2022), radiotherapy planning Cardenas et al. (2019); LaBella et al. (2024), and cardiac imaging on Myocardial Segmentation et al. (2002); Petitjean & Dacher (2011). Contemporary MIS methods Isensee et al. (2021) have substantially improved the delineation of anatomical structures and lesions. However, differences in acquisition conditions and patient populations give rise to distinct data distributions across clinical sites Yoon et al. (2024) (Fig. 1(a)), often degrading performance on unseen sites. Privacy and data-sharing constraints further limit access to images from other sites Kaissis et al. (2020). These factors pose *a key challenge to developing segmentation models that generalize to unseen clinical sites without accessing data from those sites during training* Ouyang et al. (2023).

Vision foundation models (VFMs), such as SAM and DINOv2, have attracted considerable interest for their transferable capability, offering promising opportunities for MIS. Existing methods exploit such capabilities through medical adaptation or prompt tuning Ma et al. (2024). For example, 3DSAM-Adapter Gong et al. (2024) adapts SAM from 2D to 3D through parameter-efficient modifications for promptable volumetric segmentation. SAT Liu et al. (2025) enables training-free medical segmentation through reference-guided automatic prompting. Such methods reduce the number of

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.

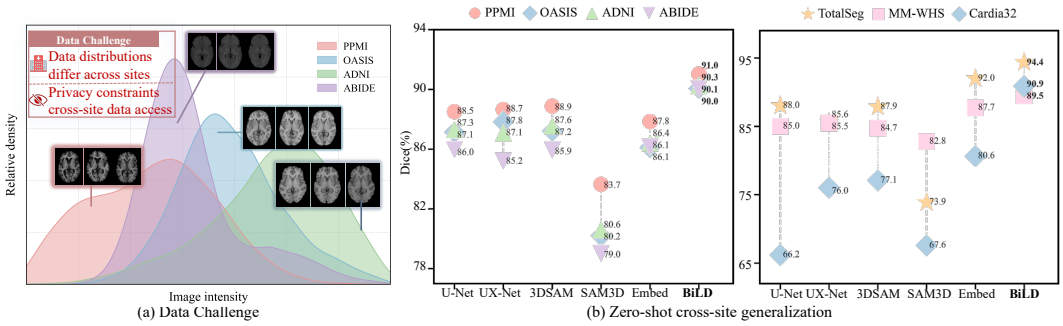


Figure 1: (a) Cross-site data distribution shifts (PPMI, OASIS, ADNI, and ABIDE) and privacy constraints on cross-site data access. (b) Zero-shot segmentation performance (Dice, %) on brain and cardiac datasets without target-site adaptation.

trainable parameters but do not reduce inference-time computation, *posing challenges for deployment in resource-constrained clinical environments*.

To reduce deployment burden, a common strategy in medical VFM distillation involves teacher adaptation (TA) and student distillation (SD) for transferring knowledge from foundation models to lightweight medical segmentation models Hinton et al. (2015); Huang et al. (2025). TA adapts a foundation-model teacher to the medical segmentation task, while SD transfers its knowledge to a lightweight segmentation student by matching the teacher's predictions or intermediate features. Existing VFMs methods alternately optimizae TA and SD using source-domain data resulting overfitting, without transfering VFM's generalization capability on unseen sites. This raises a central question:

How can we coordinate TA and SD to improve cross-site generalization without access to unseen-site data during training?

In this paper, we propose a Generalizable Bilevel Knowledge Distillation (BiLD) framework that couples SD with TA to improve cross-site generalization for 3D MIS. Specifically, the lower level performs TA for adapting a VFM teacher to the medical segmentation task, while the upper level performs SD to train a lightweight 3D segmentation student. Instead of teating TA and SD as independent tasks, we introduce a dynamic response strategy incorporating the response of TA into the upper-level SD updates, thereby coordinating teacher adaptation and knowledge transfer toward the student's cross-site generalization objective. To further facilitate the transfer of generalizable knowledge, we introduce a pseudo-site generator to construct diverse site-style variations from source-domain data, enabling the upper-level SD objective to account for simulated cross-site shifts without access to real target-site data.

We evaluate BiLD on brain and cardiac segmentation against conventional segmentation models, direct VFM adaptation methods, and VFM-based distillation methods. ① Across five unseen-domain datasets, BiLD improves segmentation performance by 4% and 6% on brain and cardiac tasks, respectively, without target-domain adaptation. ② When further fine-tuned on the target domains, BiLD achieves improvements of 2% and 4% on the respective tasks. ③ Ablation studies further reveal consistent gains in cross-domain generalization as the proposed components are progressively incorporated, supporting their individual and combined contributions.

2 The Proposed Method

2.1 Preliminaries

BiLD is a VFM-based bilevel knowledge distillation framework that involves teacher adaptation (TA) and student distillation (SD) for transferring knowledge from foundation models to lightweight medical segmentation models. The overall framework is illustrated in Fig. 2. Let $\mathcal{D}_{\text{src}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the labeled source-domain dataset, where x_i is a 3D image volume and y_i is its voxel-wise annotation. We denote the pseudo-site dataset generated by our pseudo-site generator $\mathcal{G}$ as $\mathcal{D}_{\text{ps}} = \{(\widetilde{x}_i, y_i)\}_{i=1}^{N}$, which is detailed in Sec. 2.4. The VFM teacher and the lightweight 3D student are denoted by $T_{\boldsymbol{\omega}}$ and $S_{\boldsymbol{\theta}}$, respectively.

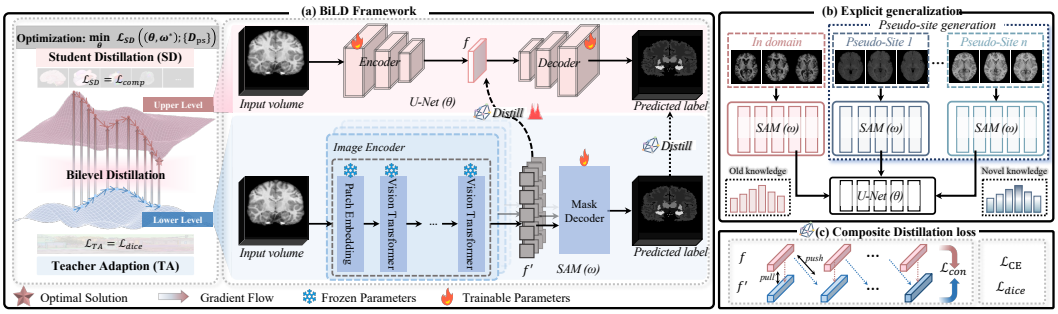


Figure 2: Generalizable Bilevel Knowledge Distillation (BiLD) framework for 3D medical image segmentation. (a) Lower-level TA adapts the teacher on $\mathcal{D}_{\text{src}}$; upper-level SD trains the student using $\mathcal{D}_{\text{ps}}$ generated from source-domain samples. The frozen SAM encoder extracts slice-wise features aggregated to 3D, followed by a trainable 3D mask decoder. (b) Source-derived pseudo-sites introduce simulated site shifts without accessing target-site data. (c) Composite loss: Dice/CE supervision plus feature-level contrastive distillation.

Teacher Configuration. BiLD is not tied to a specific VFM architecture and can be instantiated with different VFMs that provide transferable representations and support differentiable task adaptation. For clarity, we use SAM adapted for 3D medical image segmentation as the default teacher $T_{\boldsymbol{\omega}}$ to introduce the framework Kirillov et al. (2023). Specifically, the pretrained SAM image encoder extracts slice-wise representations from a 3D volume, which are aggregated into volumetric features and processed by a trainable 3D mask decoder. The image encoder remains frozen during TA, while the 3D decoder parameterized by $\boldsymbol{\omega}$ is optimized for the medical segmentation task. We further evaluate BiLD with representative VFM teachers in Sec. 3 to examine its compatibility across different teacher architectures.

2.2 Generalizable Bilevel Knowledge Distillation (BiLD)

BiLD formulates TA and SD as a coupled bilevel problem Franceschi et al. (2018). The lower-level TA objective adapts the VFM teacher $T_{\boldsymbol{\omega}}$ to the medical segmentation task using $\mathcal{D}_{\text{src}}$, while the upper-level SD objective trains the lightweight 3D student $S_{\boldsymbol{\theta}}$ using $\mathcal{D}_{\text{ps}}$. Each pseudo-site sample retains its source annotation and its corresponding source view for teacher guidance. These source-derived views allow SD to account for simulated cross-site shifts without accessing target-site images. We write the bilevel objective as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \mathcal{L}_{\text{SD}}(\boldsymbol{\theta}, \boldsymbol{\omega}^*(\boldsymbol{\theta}); \mathcal{D}_{\text{ps}}), \\ \text{s.t.} \quad & \boldsymbol{\omega}^*(\boldsymbol{\theta}) \in \arg\min_{\boldsymbol{\omega}} \mathcal{L}_{\text{TA}}(\boldsymbol{\theta}, \boldsymbol{\omega}; \mathcal{D}_{\text{src}}). \end{aligned} \tag{1}$$

Here, $\mathcal{L}_{\text{TA}}$ denotes the lower-level teacher adaptation objective, and $\mathcal{L}_{\text{SD}}$ denotes the upper-level student distillation objective. The latter is implemented by the composite objective $\mathcal{L}_{\text{comp}}$ in Eq. (8), which combines Dice and cross-entropy supervision with feature-level contrastive distillation. The pseudo-site objective serves as a surrogate for performance under site shifts. Its effectiveness on real unseen sites is assessed experimentally.

The response derivation below assumes that TA depends differentiably on $\boldsymbol{\theta}$ and that SD depends on the trainable teacher parameters $\boldsymbol{\omega}$. Without these dependencies, the corresponding response term vanishes. In particular, matching only frozen encoder features does not make SD depend on the teacher decoder.

2.3 Dynamic Response Strategy

To optimize Eq. (1), we account for the dependence of TA on the student parameters when computing SD updates. Define the reduced upper-level objective as $F(\boldsymbol{\theta}) = \mathcal{L}_{\text{SD}}(\boldsymbol{\theta}, \boldsymbol{\omega}^*(\boldsymbol{\theta}); \mathcal{D}_{\text{ps}})$. For a differentiable local solution $\boldsymbol{\omega}^*(\boldsymbol{\theta})$, its hypergradient consists of a direct term and a teacher-response term:

$$\nabla_{\boldsymbol{\theta}} F = \partial_{\boldsymbol{\theta}} \mathcal{L}_{\text{SD}} + \left(\frac{\partial \boldsymbol{\omega}^*}{\partial \boldsymbol{\theta}} \right)^{\top} \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\text{SD}}, \tag{2}$$

Algorithm 1 BiLD: schematic response-aware bilevel knowledge distillation

Require: Source training data $\mathcal{D}_{\mathrm{src}}$; pseudo-site generator $\mathcal{G}$; student $\boldsymbol{\theta}$; teacher decoder $\boldsymbol{\omega}$; total training iterations T; update budgets K, M; specified coupled TA/SD losses and a valid curvature operator $\mathbf{G}$.
Ensure: Trained student $S_{\boldsymbol{\theta}}$ for teacher-free inference.

```
Freeze the pretrained SAM image encoder; initialize the Adam optimizers.
for t = 1, ..., T do
    Sample source batches B_TA, B_SD ⊂ D_src.
    Obtain B̃_SD with G; retain paired source views and labels.
    for k = 1, ..., K do
        Update ω with Adam using ∇_ω L_TA(θ, ω; B_TA); hold θ fixed.
    end for
    Set ω̂ ← ω as the approximate lower-level solution.
    Evaluate L_SD(θ, ω̂; B̃_SD) using paired views.
    Compute g_θ = ∂_θ L_SD and g_ω = ∇_ω L_SD.
    Approximately solve G v = g_ω with at most M CG steps.
    Compute c = ∂_θ[(∇_ω L_TA)^⊤ v] at ω̂; hold ω̂, v fixed.
    Update θ with Adam using g_θ − c; hold ω̂ fixed.
end for
return S_θ; discard the teacher for inference.
```

where partial derivatives hold the other parameter block fixed. Gradients are column vectors, and all derivatives are evaluated at $(\boldsymbol{\theta}, \boldsymbol{\omega}^*(\boldsymbol{\theta}))$ unless stated otherwise. The response term accounts for the indirect effect of TA on the SD objective evaluated using pseudo-sites, as illustrated in Fig. 3. At a stationary lower-level solution, $\nabla_{\boldsymbol{\omega}} \mathcal{L}_{\mathrm{TA}}(\boldsymbol{\theta}, \boldsymbol{\omega}^*; \mathcal{D}_{\mathrm{src}}) = \mathbf{0}$. Define $\mathbf{H} = \partial(\nabla_{\boldsymbol{\omega}} \mathcal{L}_{\mathrm{TA}})/\partial \boldsymbol{\omega}$ and $\mathbf{C} = \partial(\nabla_{\boldsymbol{\omega}} \mathcal{L}_{\mathrm{TA}})/\partial \boldsymbol{\theta}$. For $\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}$ and $\boldsymbol{\omega} \in \mathbb{R}^{d_\omega}$, $\mathbf{H} \in \mathbb{R}^{d_\omega \times d_\omega}$ and $\mathbf{C} \in \mathbb{R}^{d_\omega \times d_\theta}$. Assuming sufficient smoothness and a nonsingular $\mathbf{H}$, implicit differentiationLorraine et al. (2020); Rajeswaran et al. (2019) gives

$$\frac{\partial \boldsymbol{\omega}^*}{\partial \boldsymbol{\theta}} = -\mathbf{H}^{-1}\mathbf{C}. \tag{3}$$

We approximate the inverse-curvature action using a Gauss–Newton matrix $\mathbf{G}$. For a least-squares residual representation $\mathcal{L}_{\mathrm{TA}} = \frac{1}{2}\|\mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\omega}; \mathcal{D}_{\mathrm{src}})\|_2^2$, this approximation takes the form

$$\mathbf{H} \approx \mathbf{G} := \mathbf{J}^\top \mathbf{J}, \qquad \mathbf{J} = \frac{\partial \mathbf{r}}{\partial \boldsymbol{\omega}}. \tag{4}$$

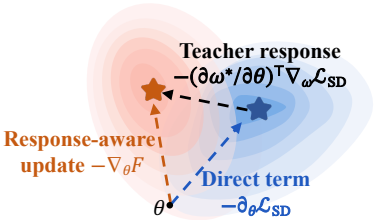


Figure 3: **Dynamic response strategy.** Conceptual illustration of Eq. 2, showing how the teacher-response term corrects the direct SD update direction.

For other TA losses, the curvature approximation must be specified consistently with the chosen loss. The matrix $\mathbf{G}$ is distinct from the pseudo-site generator $\mathcal{G}$. Matrix–vector products avoid explicitly forming and inverting the full Hessian; the mixed derivative in the response term is still required.

Given a positive-definite curvature system, we compute a response vector $\mathbf{v}$ by approximately solving

$$\mathbf{G}\mathbf{v} = \nabla_{\boldsymbol{\omega}} \mathcal{L}_{\mathrm{SD}}, \tag{5}$$

using Gauss–Newton-vector products and a limited number of conjugate-gradient (CG) iterations. For the residual form in Eq. (4), a product is computed as $\mathbf{G}\mathbf{u} = \mathbf{J}^\top(\mathbf{J}\mathbf{u})$. A singular curvature matrix requires an explicitly specified regularization or an alternative solver. The resulting approximate student hypergradient is

$$\nabla_{\boldsymbol{\theta}} F \approx \partial_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{SD}} - \mathbf{C}^\top \mathbf{v}. \tag{6}$$

Eq. (6) incorporates the TA response into student updates for the pseudo-site SD objective. Its role is to coordinate the coupled objectives; performance on simulated shifts remains a surrogate for generalization to real unseen sites. In practice, K teacher-decoder updates yield an approximate lower-level solution $\widehat{\boldsymbol{\omega}}$, at which the derivatives are evaluated before updating $\boldsymbol{\theta}$. Algorithm 1 summarizes this procedure under the stated coupling and curvature assumptions.

2.4 Pseudo-Site Generator

To support SD under simulated site shifts without target-site data, we construct pseudo-sites from labeled source volumes:

$$\widetilde{x}_i = \mathcal{G}(x_i; \xi_i), \qquad \widetilde{y}_i = y_i, \tag{7}$$

where ξ_i denotes randomly sampled transformation settings. We use three acquisition-motivated profiles: (i) intensity and contrast changes with smooth bias fields; (ii) Gaussian noise with optional 3D smoothing; and (iii) directional averaging to approximate reduced spatial detail. Settings are sampled per volume rather than per slice. These transformations alter appearance without additional spatial warping, allowing the source annotation to be reused.

Each epoch, a randomly selected subset of source volumes receives additional transformations, while the remaining volumes retain their source views. Within BiLD, TA uses source-domain data. For SD, the teacher encoder processes the paired source view, whereas the student processes its pseudo-site counterpart. The shared labels and paired features support segmentation supervision and contrastive distillation, respectively. This construction exposes the upper-level objective to simulated appearance shifts without requiring images or statistics from unseen sites.

2.5 Loss Functions

We distinguish the upper-level student distillation (SD) objective from the source-supervised loss used for lower-level teacher adaptation (TA).

Upper-Level Loss. The SD objective combines segmentation supervision on pseudo-site views with contrastive distillation on pseudo-site data $\mathcal{D}_{\text{ps}}$:

$$\mathcal{L}_{\text{comp}} = \alpha\, \mathcal{L}_{\text{Dice}}(p, y) + \beta\, \mathcal{L}_{\text{CE}}(p, y) + \gamma\, \mathcal{L}_{\text{Con}}(E_S, E_T), \tag{8}$$

where p is the student prediction, y is the source annotation, and α, β, γ are loss weights. E_S and E_T denote student features from the pseudo-site view and frozen teacher encoder features from its paired source view, respectively. The upper-level objective $\mathcal{L}_{\text{SD}}$ averages $\mathcal{L}_{\text{comp}}$ over paired samples.

Lower-Level Loss. TA adapts the teacher on source-domain data $\mathcal{D}_{\text{src}}$ using Dice loss:

$$\mathcal{L}_{\text{TA}}(\boldsymbol{\omega}; \mathcal{D}_{\text{src}}) = \frac{1}{|\mathcal{D}_{\text{src}}|} \sum_{(x,y)\in\mathcal{D}_{\text{src}}} \mathcal{L}_{\text{Dice}}(q_{\boldsymbol{\omega}}(x), y)\,, \tag{9}$$

where $q_{\boldsymbol{\omega}}(x)$ denotes the teacher's voxel-wise class probabilities, obtained from the frozen SAM encoder and the trainable 3D decoder. Minimizing this term updates the decoder parameters $\boldsymbol{\omega}$ while keeping the encoder fixed. This source-supervised term is independent of the student parameters $\boldsymbol{\theta}$; it does not, by itself, define the differentiable TA–SD coupling required by the response formulation in Sec. 2.3.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate our method on two representative medical segmentation benchmarks, including a brain dataset and a cardiac dataset. For brain segmentation, we use PPMI Marcus et al. (2010) as the primary training dataset, which contains 13 anatomical regions plus background. To evaluate cross-site generalization, additional brain datasets from different sites, including OASIS Marcus et al. (2010), ABIDE Di Martino et al. (2014), and ADNI Di Martino et al. (2014), are introduced. For cardiac segmentation, we use 200 3D cardiac CT images selected from the TotalSeg (Wasserthal et al., 2023) as the primary training set, with 4 target cardiac structures and background. Additional cardiac datasets from different sites, including MM-WHS Zhuang et al. (2019) and CAT08 Schaap et al. (2009) are introduced to evaluate generalization.

Evaluation Metrics. We evaluate segmentation performance using the Dice similarity coefficient (DSC) and the 95th-percentile Hausdorff distance (HD95). DSC measures the volumetric overlap between predicted and ground-truth segmentations, while HD95 quantifies boundary discrepancies. Higher DSC and lower HD95 indicate better segmentation performance.

Implementation Details. Our framework uses SAM with a frozen encoder and a trainable lightweight 3D mask decoder as the lower-level teacher, and trains a full 3D U-Net as the upper-level student. All models are optimized using the Adam optimizer with default hyperparameters, a learning rate of 1×10^{-4}, and a batch size of 2. All experiments are conducted on a single NVIDIA A40 (48GB) GPU. All volumes are resampled to $128 \times 128 \times 128$. The loss weights are set to

Table 1: **Cross-domain segmentation comparison.** Methods are trained on PPMI and TotalSeg for brain and cardiac segmentation, respectively. Dice scores (%, higher is better) are reported as mean ± standard deviation across test cases. Avg. denotes the mean Dice across out-of-domain datasets. The best result in each column is highlighted in **bold**.

Method	*Brain: Train on PPMI*					*Cardiac: Train on TotalSeg*			
	PPMI	OASIS	ADNI	ABIDE	Avg.	TotalSeg	MM-WHS	CAT08	Avg.
Traditional methods									
U-Net	88.50 $\pm$1.42	87.13 $\pm$0.78	87.32 $\pm$2.13	85.96 $\pm$1.02	86.80	88.01 $\pm$9.62	84.98 $\pm$6.34	66.18 $\pm$15.59	75.58
3D UX-Net	88.65 $\pm$1.55	87.83 $\pm$0.80	87.11 $\pm$2.57	85.20 $\pm$1.28	86.71	85.56 $\pm$8.73	85.46 $\pm$6.78	76.03 $\pm$10.66	80.75
SwinUNETR	88.61 $\pm$1.36	87.39 $\pm$0.71	87.11 $\pm$2.25	85.21 $\pm$1.44	86.57	86.98 $\pm$10.01	80.16 $\pm$10.35	83.76 $\pm$6.20	81.96
Direct adaptation methods									
3DSAM-Adapter	88.88 $\pm$1.21	87.21 $\pm$0.74	87.58 $\pm$1.80	85.92 $\pm$1.17	86.90	87.86 $\pm$6.77	84.69 $\pm$7.19	77.08 $\pm$8.79	80.89
SAM3D	83.65 $\pm$1.40	80.20 $\pm$1.08	80.62 $\pm$2.54	79.00 $\pm$1.82	79.94	73.89 $\pm$20.18	82.76 $\pm$5.13	67.57 $\pm$11.10	75.16
MaskSAM	88.39 $\pm$1.26	86.49 $\pm$1.06	86.79 $\pm$2.23	85.32 $\pm$1.47	86.20	73.35 $\pm$11.16	56.59 $\pm$22.16	67.04 $\pm$13.85	61.81
VFMs knowledge distillation methods									
KnowSAM	73.13 $\pm$2.68	71.24 $\pm$3.05	68.84 $\pm$7.57	65.30 $\pm$6.18	68.46	61.96 $\pm$13.51	74.56 $\pm$4.74	53.44 $\pm$12.97	64.00
ESP-MedSAM	86.71 $\pm$1.34	84.44 $\pm$1.05	85.44 $\pm$2.20	82.54 $\pm$2.77	84.14	77.83 $\pm$9.93	75.08 $\pm$7.54	70.30 $\pm$11.45	72.69
IAD	67.76 $\pm$1.74	67.59 $\pm$1.87	65.49 $\pm$3.42	57.28 $\pm$5.19	63.45	79.57 $\pm$9.00	66.11 $\pm$10.49	39.52 $\pm$7.01	52.82
Embed-MedSAM	87.85 $\pm$1.59	86.10 $\pm$0.90	86.42 $\pm$2.24	86.12 $\pm$1.67	86.21	91.99 $\pm$5.67	87.70 $\pm$2.45	80.64 $\pm$10.31	84.17
BiLD (ours)	**91.03** $\pm$1.43	**90.06** $\pm$1.78	**90.27** $\pm$0.88	**90.03** $\pm$0.59	**90.13**	**94.38** $\pm$3.51	**89.45** $\pm$3.13	**90.87** $\pm$5.58	**90.16**

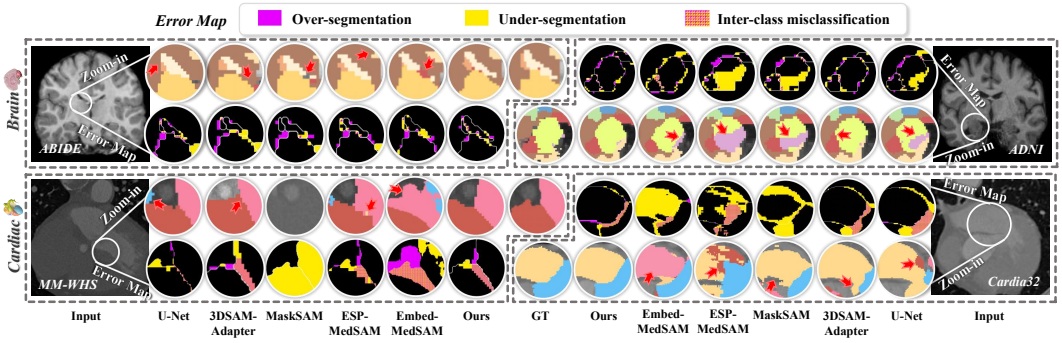


Figure 4: **Qualitative segmentation comparisons on out-of-domain datasets.** In the error maps, purple and indicates over-segmentation, yellow indicates under-segmentation and white contours denote GT boundaries.

$\alpha = 10$, $\beta = 1$, and $\gamma = 1$, with a contrastive temperature of $\tau = 0.1$. During training, we apply standard 3D spatial augmentations and use the pseudo-site generator (PSG) to generate additional training samples for improved generalization.

3.2 Experimental Results

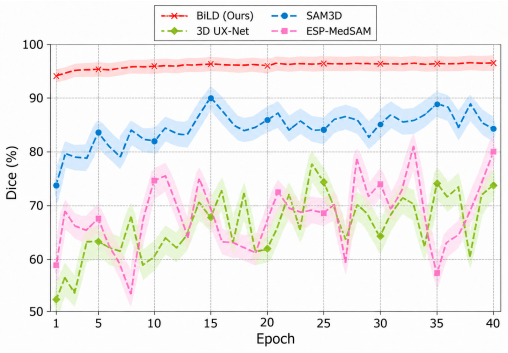
We compare source-domain and out-of-domain Dice scores for brain and cardiac segmentation, grouping baselines into traditional segmentation, direct SAM adaptation, and SAM knowledge distillation methods. The distillation baselines include KnowSAM Huang et al. (2025), ESP-MedSAM Xu et al. (2024), IAD Hong et al. (2026), and Embed-MedSAM Zhang et al. (2025). For 2D adaptation and distillation methods, slice-wise predictions are stacked into 3D segmentation volumes. Prompt-based baselines are evaluated using prompts automatically generated from ground-truth masks. Embed-MedSAM denotes our RepViT-based adaptation for automatic segmentation, rather than a reproduction of the official implementation.

Quantitative Comparison. As shown in Table 1, BiLD achieves the highest Dice scores on the source-domain and all out-of-domain test datasets. Its mean out-of-domain Dice reaches 90.13% for brain segmentation and 90.16% for cardiac segmentation, surpassing the strongest baseline in each task by 3.23 and 5.99 percentage points, respectively. Notably, several baselines exhibit substantial performance degradation on CAT08, whereas BiLD maintains a Dice score of 90.87%, highlighting its cross-domain generalization capability.

Qualitative Comparison. Fig. 4 shows that BiLD produces more accurate segmentations with closer alignment to ground-truth boundaries and reduced over- and under-segmentation on out-of-

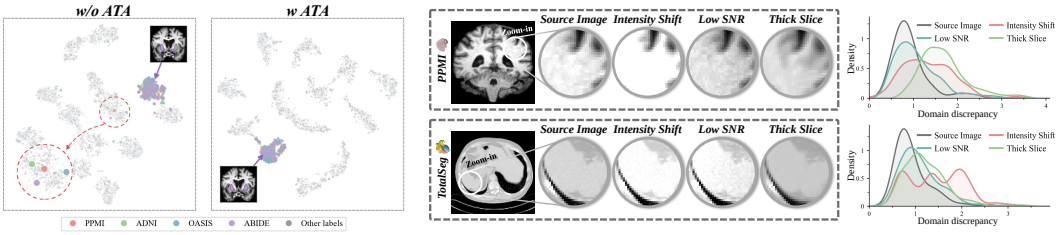
(a) Fine-tuning results.

Method	*Brain* PPMI → ADNI		*Cardiac* TotalSeg → CAT08	
	Dice↑	HD95↓	Dice↑	HD95↓
Traditional methods				
U-Net	88.76 ± 0.63	1.36 ± 0.13	89.09 ± 9.06	14.63 ± 13.34
3D UX-Net	89.01 ± 0.60	1.32 ± 0.13	87.37 ± 6.24	25.08 ± 11.80
SwinUNETR	88.96 ± 0.63	1.32 ± 0.12	88.09 ± 6.71	11.91 ± 7.27
Direct adaptation methods				
3DSAM-Adapter	88.88 ± 0.59	1.81 ± 0.15	91.67 ± 4.03	4.21 ± 2.27
SAM3D	81.95 ± 1.50	3.01 ± 0.35	88.99 ± 4.95	8.90 ± 9.06
MaskSAM	87.72 ± 1.00	1.88 ± 0.14	74.02 ± 12.60	21.23 ± 9.38
SAM knowledge distillation methods				
KnowSAM	73.07 ± 2.05	4.76 ± 0.76	80.29 ± 10.75	22.00 ± 5.45
ESP-MedSAM	86.38 ± 0.97	2.04 ± 0.16	81.35 ± 13.16	20.17 ± 7.99
IAD	70.33 ± 1.94	4.26 ± 0.32	78.39 ± 4.49	34.25 ± 5.51
Embed-MedSAM	87.46 ± 0.72	1.59 ± 0.15	89.07 ± 8.16	9.03 ± 9.35
BiLD (ours)	**91.26** ± 0.61	**1.18** ± 0.07	**95.04** ± 3.72	**2.30** ± 1.79



(b) Cardiac fine-tuning curves.

Figure 5: **Target-domain fine-tuning.** (a) Results on ADNI and CAT08, reported as mean ± standard deviation of Dice (%) and HD95 (mm). Best results are highlighted in bold. (b) Dice curves for TotalSeg → CAT08.



(a) Cross-dataset features

(b) Pseudo-site generation

Figure 6: **Cross-dataset features and pseudo-site generation.** (a) Colors distinguish datasets for the same class; gray denotes other classes. (b) Augmented images and domain-discrepancy distributions for brain and cardiac data.

domain samples, further supporting its cross-domain generalization capability. The zoomed-in regions reveal that competing methods sometimes miss substantial portions of the target or extend into surrounding tissues, whereas BiLD more faithfully preserves the target morphology.

Fine-Tuning Evaluation. We further fine-tune the source-trained methods on ADNI and CAT08. As shown in Fig. 5 (a), BiLD achieves the highest Dice and lowest HD95 on both datasets, retaining its segmentation accuracy and boundary precision. Fig. 5 (b) further shows higher Dice scores with smaller fluctuations during fine-tuning. These results suggest that BiLD learns representations that not only generalize directly to new datasets but also provide an effective initialization for subsequent adaptation.

3.3 Mechanistic Analysis of BiLD

Why can we distill generalization?

To understand why generalization can be distilled, we examine the roles of alternating teacher adaptation (ATA), the pseudo-site generator (PSG), and bilevel knowledge distillation, respectively.

ATA combines teacher adaptation with the composite distillation loss to transfer teacher knowledge to the student. As shown in Fig. 6, ATA brings same-class student features from different datasets closer together and promotes cross-domain mixing, suggesting reduced sensitivity to dataset-specific variations. This observation supports the contribution of teacher guidance to cross-domain representation learning beyond source-domain fitting.

PSG introduces complementary diversity through intensity and bias-field perturbations, low-SNR corruption, and single-axis blurring.

Table 2: **Ablation study on the source domain and three unseen domains.** All models are trained and selected using source-domain data only. The best results are highlighted in bold.

Method	*Brain: Train on PPMI*							
	PPMI		OASIS		ADNI		ABIDE	
	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓
Supervised U-Net	88.66	1.23	86.39	1.51	86.83	1.52	86.88	1.41
+ ATA	88.80	1.27	86.44	1.49	87.02	1.52	87.28	1.43
+ PSG	88.79	1.24	86.90	1.47	87.34	1.42	86.84	1.42
+ **BiLD (ours)**	**91.03**	**1.20**	**90.06**	**1.33**	**90.27**	**1.27**	**90.03**	**1.38**

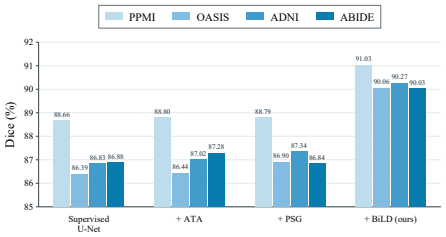


Figure 7: **Cross-domain comparison.** Dice (%) of different ablation configurations on brain datasets.

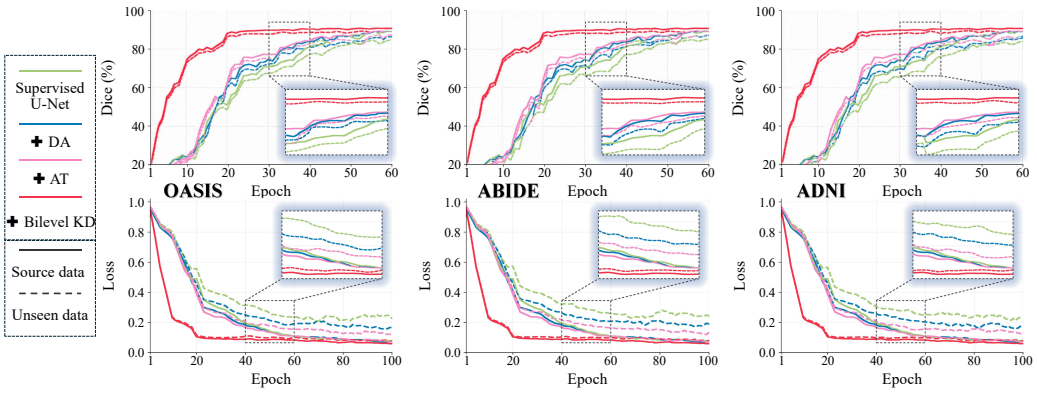


Figure 8: **Training and cross-domain validation curves.** Dice scores (top) and loss values (bottom) for different ablation configurations. The gap between training and validation curves reflects the generalization gap across domains.

As shown in Fig. 6, these transformations introduce diversity in imaging appearance and data distribution while preserving anatomical semantics. This encourages the student to learn stable semantic representations instead of relying on source-specific imaging characteristics.

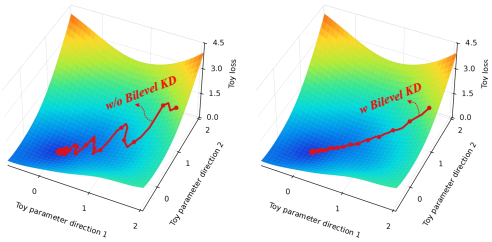


Figure 9: **Gradient-based update trajectories with and without Bilevel Knowledge Distillation.**

Beyond output and feature matching, **Bilevel knowledge distillation** incorporates the response of teacher adaptation into student updates to coordinate the two objectives at the gradient level. Fig. 9 compares student gradients with and without response correction to examine their alignment with gradients evaluated on held-out pseudo-site samples. The curves in Fig. 8 further show that adding Bilevel Knowledge Distillation leads to faster, more stable convergence and better out-of-domain performance.

With these two mechanisms, we don't just distill knowledge. We distill generalization from VFMs.

A Plug-and-Play Distillation Framework. We evaluate BiLD with different VFMs by replacing either the entire teacher or its pretrained encoder while retaining a shared 3D decoder. Table 3 shows BiLD has a strong out-of-domain segmentation performance across multiple teacher configurations, indicating which is not tied to a specific teacher architecture. These results support BiLD as a general, plug-and-play distillation framework for transferring knowledge from diverse VFMs to lightweight 3D segmentation models for cross-domain generalization. As shown in Fig. 6, BiLD brings features of same-class from different datasets closer together than U-Net, suggesting stronger generalization under domain shifts.

Table 3: **Teacher configuration ablation.** We replace either the full teacher or its encoder. Student cardiac Dice (%, higher is better) is reported. Avg. denotes the mean Dice across out-of-domain datasets.

Teacher configuration	TotalSeg	MM-WHS	CAT08	Avg.
A. Full teacher replacement				
3DSAM-Adapter	92.65	87.63	90.04	88.83
ProtoSAM-based teacher*	92.35	90.17	88.12	89.04
SAT (w/o text conditioning)	92.05	90.21	90.32	90.27
B. Encoder replacement (shared 3D decoder)				
DINOv2 encoder	94.25	90.63	91.69	91.16
MedSAM encoder	94.43	90.96	92.07	91.52
BiomedParse encoder	92.14	88.91	89.09	89.00
MARS encoder	94.06	91.41	90.51	90.96
SAM encoder (ours)	94.38	89.45	90.87	90.16

Table 4: Contrastive distillation ablation.

Feature objective	OOD DSC↑	HD95↓
None ($\lambda = 0$)	87.84	1.48
MSE matching	88.57	1.59
Contrastive (ours)	88.88	1.45

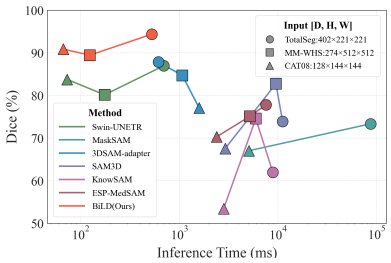


Figure 10: Performance–efficiency trade-off.

Ablation Studies. Beyond evaluating generalization performance of BiLD, we conduct ablation studies to examine the contribution of individual components. Table 2 and Fig. 7 show that the full BiLD framework achieves higher Dice scores while maintaining consistent performance across out-of-domain datasets. Fig. 8 also shows a smaller gap between source-domain training and out-of-domain validation performance, alongside strong validation results, supporting its cross-domain generalization c.

Additionally, we conduct an ablation study on the contrastive distillation objective in Table 4 ,which shows that MSE matching underperforms no feature distillation, suggesting that an unsuitable objective can hinder student learning, possibly by imposing overly restrictive element-wise alignment. In contrast, contrastive distillation provides relation-based flexible supervision and improves both metrics.

3.4 Performance–Efficiency Trade-off.

Beyond the generalization gains, we also examine the performance–efficiency trade-off, as shown in Fig. 10. We compared segmentation accuracy and inference time for small (CAT08), medium (Totalseg), and large (MM-WHS) inputs. In contrast, most competing methods incur higher inference costs without corresponding accuracy gains, whereas BiLD achieves higher Dice scores with shorter inference times across all three input scales. However, BiLD maintains high accuracy with low inference latency across all three input scales, demonstrating a favorable performance–efficiency trade-off.

4 Conclusion

In this paper, we presented BiLD, a generalization-oriented bilevel distillation framework that couples SAM adaptation with student learning for robust medical image segmentation under domain shifts. The lower level adapts SAM into a medically competent teacher, while the upper level trains a lightweight, prompt-free 3D student through generalization-driven distillation across domains. BiLD employs an efficient dynamic response strategy and a composite distillation objective that jointly promotes region consistency, boundary fidelity, and representation alignment.Extensive experiments across brain and cardiac benchmarks demonstrate improved segmentation accuracy, cross-dataset generalization, and inference efficiency. Beyond SAM-based medical segmentation, the bilevel formulation offers a general paradigm for jointly optimizing foundation-model adaptation and knowledge transfer, rather than treating the teacher as a fixed source of supervision. *More broadly*, BiLD provides a general distillation paradigm that can potentially transfer transferable priors from other foundation models to diverse lightweight architectures and downstream tasks.

AI USE STATEMENT

During the preparation of this work, we used generative AI tools to assist with English translation, language polishing, and LaTeX formatting. Generative AI tools were also used to assist in drafting and rephrasing portions of the manuscript. All AI-assisted text was carefully reviewed, verified, and revised by the authors. The experimental design, method development, implementation, experiments, analysis, and conclusions were conducted and validated by the authors. Generative AI outputs were not used as sources of factual evidence or scientific references, and all cited references were checked by the authors. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work studies cross-site generalization for 3D medical image segmentation using existing research datasets and does not involve the collection of new patient data, interaction with human subjects, or clinical intervention. All datasets are used in accordance with their respective access conditions and usage requirements. We do not attempt to identify or recover personal information from the medical images. Although BiLD is designed to improve robustness under cross-site distribution shifts, its performance may still vary across imaging devices, acquisition protocols, and patient populations. The proposed method is intended for research purposes and should not be used directly for clinical decision making without appropriate external validation, regulatory review, and human oversight.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed BiLD framework, including the bilevel formulation, dynamic response strategy, pseudo-site generator, loss functions, and training procedure in the main paper. The experimental setup specifies the datasets, evaluation metrics, preprocessing, model configurations, optimization settings, and major hyperparameters used in our experiments. The appendix provides additional methodological background and complementary evaluations, including HD95 results, case-level performance distributions, and qualitative analyses of different teacher configurations. Together, these details are intended to support reproduction and independent evaluation of the proposed method.

REFERENCES

Carlos E Cardenas, Jinzhong Yang, Brian M Anderson, Laurence E Court, and Kristy B Brock. Advances in auto-segmentation. In *Seminars in radiation oncology*, volume 29, pp. 185–197. Elsevier, 2019.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

Mohamed S Elmahdy, Laurens Beljaards, Sahar Yousefi, Hessam Sokooti, Fons Verbeek, Uulke A Van Der Heide, and Marius Staring. Joint registration and segmentation via multi-task learning for adaptive radiotherapy of prostate cancer. *IEEE Access*, 9:95551–95568, 2021.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jiankang Hong, Ye Luo, Yinan Liu, and Junsong Yuan. From infusion to assimilation distillation for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20985–20995, 2026.

Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, Chen Gong, and Dong Liang. Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature machine intelligence*, 2 (6):305–311, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Dominic LaBella, Katherine Schumacher, Michael Mix, Kevin Leu, Shan McBurney-Lin, Pierre Nedelec, Javier Villanueva-Meyer, Jonathan Shapey, Tom Vercauteren, Kazumi Chia, et al. Brain tumor segmentation (brats) challenge 2024: Meningioma radiotherapy planning automated segmentation. *arXiv e-prints*, pp. arXiv–2405, 2024.

Xueyu Liu, Guangze Shi, Rui Wang, Yexin Lai, Jianan Zhang, Weixia Han, Min Lei, Ming Li, Xiaoshuang Zhou, Yongfei Wu, Chen Wang, and Wen Zheng. Segment any tissue: One-shot reference guided training-free automatic point prompting for medical image segmentation. *Medical Image Analysis*, 102:103550, 2025. doi: 10.1016/j.media.2025.103550.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1540–1552. PMLR, 2020. URL https://proceedings.mlr.press/v108/lorraine20a.html.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010.

American Heart Association Writing Group on Myocardial Segmentation, Registration for Cardiac Imaging:, Manuel D Cerqueira, Neil J Weissman, Vasken Dilsizian, Alice K Jacobs, Sanjiv Kaul, Warren K Laskey, Dudley J Pennell, John A Rumberger, Thomas Ryan, et al. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. *Circulation*, 105(4):539–542, 2002.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2023. doi: 10.1109/TMI.2022.3224067. URL https://doi.org/10.1109/TMI.2022.3224067.

Caroline Petitjean and Jean-Nicolas Dacher. A review of segmentation methods in short axis cardiac mr images. *Medical image analysis*, 15(2):169–184, 2011.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html.

Michiel Schaap, Coert T Metz, Theo van Walsum, Alina G van der Giessen, Annick C Weustink, Nico R Mollet, Christian Bauer, Hrvoje Bogunović, Carlos Castro, Xiang Deng, et al. Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. *Medical image analysis*, 13(5):701–714, 2009.

Toufique A Soomro, Lihong Zheng, Ahmed J Afifi, Ahmed Ali, Shafiullah Soomro, Ming Yin, and Junbin Gao. Image segmentation for mr brain tumor detection using machine learning: a review. *IEEE Reviews in Biomedical Engineering*, 16:70–90, 2022.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Qing Xu, Jiaxuan Li, Xiangjian He, Ziyu Liu, Zhen Chen, Wenting Duan, Chenxin Li, Maggie M He, Fiseha B Tesema, Wooi P Cheah, et al. Esp-medsam: Efficient self-prompting sam for universal domain-generalized medical image segmentation. *arXiv preprint arXiv:2407.14153*, 2024.

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, 2024.

Ying Zhang, Fangdie Ye, Xiaofeng Yu, Xin Lian, Tianlong Jiang, Liyu Yang, and Lei Yang. Embedded framework for clinical medical image segment anything in resource limited healthcare regions. *NPJ Digital Medicine*, 8(1):568, 2025.

Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019.

A APPENDIX

This appendix provides additional related work, HD95 results for brain and cardiac segmentation, case-level performance distributions, and qualitative comparisons on source domains and across teacher configurations.

A.1 RELATED WORK

A.1.1 MEDICAL IMAGE SEGMENTATION

Medical image segmentation has progressed through architectures that combine local anatomical detail with broader spatial context. U-Net (Ronneberger et al., 2015) establishes an encoder–decoder design with skip connections, while 3D U-Net (Çiçek et al., 2016) extends this design to volumetric data. SwinUNETR (Hatamizadeh et al., 2021) incorporates hierarchical self-attention, whereas 3D UX-Net (Lee et al., 2023) uses large-kernel convolutions to model volumetric context. These approaches provide strong task-specific segmentation models, but architectural improvements alone do not resolve distribution shifts between training and deployment sites. In BiLD, a compact 3D segmentation network instead serves as the student, allowing us to investigate how training objectives and foundation-model knowledge can improve its cross-site transfer.

Domain generalization addresses these shifts without using target-domain data during training. For medical imaging, differences in acquisition conditions can change image appearance while leaving the anatomical segmentation task unchanged. SynthSeg (Billot et al., 2023) exploits this observation by training on label-conditioned synthetic images with randomized contrast and resolution. The causality-inspired approach of Ouyang et al. (2023) combines global intensity nonlinear augmentation with interventional pseudo-correlation augmentation to reduce reliance on domain-specific appearance cues. These studies motivate source-derived variations as a practical route to robustness when additional sites are unavailable. Our pseudo-site generator follows this broad principle but uses label-preserving appearance perturbations of source volumes, rather than synthesizing new anatomy. Its role is to construct surrogate environments for the upper-level student distillation objective, connecting appearance diversification to the bilevel learning formulation.

Vision foundation models offer another route to transferable segmentation representations. SAM (Kirillov et al., 2023) learns promptable segmentation from large-scale natural-image data, motivating its adaptation to medical tasks. MedSAM (Ma et al., 2024) adapts SAM using medical images, while 3DSAM-Adapter (Gong et al., 2024) modifies the image encoder, prompt encoder, and mask decoder for volumetric tumor segmentation with parameter-efficient adaptation. Such approaches address the mismatch between natural-image pretraining and medical inputs, but retain a foundation-model-based segmentation pipeline at inference. Reducing the number of updated parameters is therefore distinct from reducing the computational and memory costs of the deployed model. Knowledge distillation offers a complementary strategy by transferring teacher representations or predictions to a smaller model.

Medical SAM distillation methods differ in their supervision and transfer mechanisms. KnowSAM (Huang et al., 2025) combines learnable prompting with SAM-induced distillation for semi-supervised segmentation. ESP-MedSAM (Xu et al., 2024) separates common image knowledge from modality-specific medical knowledge in a lightweight, self-prompting model and evaluates transfer beyond its training distributions. Outside medical imaging, GKD (Lv et al., 2026) explicitly targets generalizable VFM distillation by separating representation distillation from task learning and freezing the distilled representations during task adaptation. These methods establish generalization-aware distillation as a research direction beyond model compression. BiLD investigates a different formulation for fully supervised 3D medical segmentation: lower-level teacher adaptation (TA) and upper-level student distillation (SD) are linked through a teacher-response term, with source-derived pseudo-sites defining the upper-level evaluation environments. The focus is on this coupling and its effect on unseen-site performance, rather than compression alone.

A.1.2 BILEVEL OPTIMIZATION LEARNING

Bilevel learning expresses an outer objective in terms of the solution to an inner optimization problem. This structure allows a training process to be evaluated against a criterion that differs from

its immediate fitting objective. Franceschi et al. (2018) provide a unified formulation for gradient-based hyperparameter optimization and meta-learning, explicitly accounting for inner optimization dynamics. The distinction from ordinary alternating optimization is not merely the order of parameter updates: the outer update must account for how the inner solution changes with the outer variables. Alternately minimizing two losses while ignoring this dependence does not, in general, optimize the corresponding reduced outer objective.

Computing this dependence efficiently is a central challenge. Unrolled differentiation follows a finite sequence of inner updates, whereas implicit differentiation uses the stationarity conditions of the inner problem under suitable regularity assumptions. Rajeswaran et al. (2019) use implicit gradients for meta-learning without retaining the complete inner optimization trajectory. Lorraine et al. (2020) further show how approximate inverse-Hessian operations enable large-scale hyperparameter optimization, including learning data augmentation for validation performance. These methods establish tools for computing response-aware updates; they do not imply that an arbitrary outer objective will generalize to unseen domains. The outer data distribution and objective remain essential design choices.

Bilevel formulations also support adaptation and interaction between learning components. BiLoRA (Qiang et al., 2024) assigns the pseudo singular vectors and values of low-rank updates to different optimization levels and data subsets to alleviate overfitting during foundation-model adaptation. In teacher–student learning, Meta Pseudo Labels (Pham et al., 2021) updates a teacher using feedback from the performance of a student trained on its pseudo-labels, demonstrating that teacher quality can be assessed through its downstream learning effect. For medical segmentation, Fan et al. (2024) use bilevel learning to coordinate task-specific registration and segmentation decoders. These examples motivate coupled objectives, but address different variables, supervision settings, and deployment goals from VFM-to-student cross-site distillation.

BiLD places medical TA at the lower level and the compact student's SD objective at the upper level. Its dynamic response strategy uses a Gauss–Newton approximation and conjugate-gradient iterations to approximate the teacher-response contribution to the student update. Unlike validation-driven augmentation learning, our pseudo-site generator is a stochastic, non-learned transformation mechanism: it supplies source-derived environments for SD rather than introducing another set of optimized generator parameters. The response computation links the two optimization levels, while pseudo-sites make the upper objective sensitive to simulated appearance shifts without target-site training data. This constitutes a surrogate generalization objective, whose transfer to real unseen sites is evaluated empirically rather than guaranteed by the bilevel formulation itself.

A.2 In-Domain Qualitative Comparisons

Figure 11 compares source-domain predictions on PPMI and TotalSeg. Magnified overlays and error maps complement the aggregate results by showing local boundary and inter-class errors.

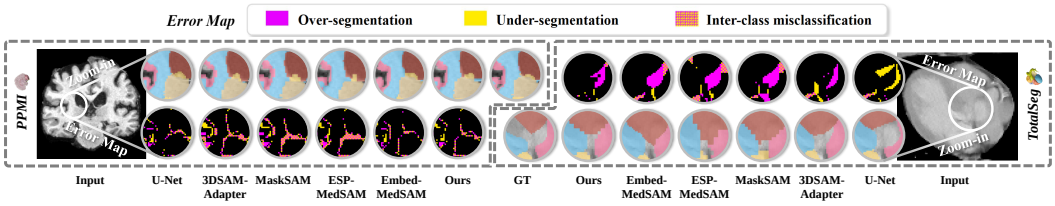


Figure 11: **In-domain segmentation on PPMI (left) and TotalSeg (right).** GT denotes ground truth; Ours denotes BiLD. Magenta, yellow, and the patterned legend entry denote over-segmentation, under-segmentation, and inter-class misclassification, respectively.

A.3 HD95 Results

Table 5 complements the Dice comparison in Table 1 of the main paper with the 95th-percentile Hausdorff distance (HD95). Lower HD95 values indicate better boundary agreement.

Table 5: **Cross-domain segmentation comparison in HD95.** Methods are trained on PPMI and TotalSeg for brain and cardiac segmentation, respectively. Lower is better. Avg. denotes the mean across out-of-domain datasets. BiLD results are highlighted in **bold**.

Method	*Brain: Train on PPMI*					*Cardiac: Train on TotalSeg*			
	PPMI	OASIS	ADNI	ABIDE	Avg.	TotalSeg	MM-WHS	CAT08	Avg.
Traditional methods									
U-Net	2.01 $\pm$0.94	1.97 $\pm$0.16	1.95 $\pm$0.51	5.68 $\pm$0.75	3.20	166.52 $\pm$118.72	94.22 $\pm$15.72	33.32 $\pm$8.95	63.77
3D UX-Net	1.71 $\pm$0.29	1.99 $\pm$0.38	2.06 $\pm$0.86	7.25 $\pm$0.31	3.77	152.74 $\pm$117.46	94.89 $\pm$16.98	37.85 $\pm$7.90	66.37
SwinUNETR	1.80 $\pm$0.25	2.05 $\pm$0.24	2.15 $\pm$1.11	7.03 $\pm$0.47	3.74	131.80 $\pm$110.51	97.72 $\pm$26.66	21.11 $\pm$8.12	59.42
Direct adaptation methods									
3DSAM-Adapter	1.84 $\pm$0.23	**1.33** $\pm$0.14	**1.32** $\pm$0.22	4.82 $\pm$0.25	2.49	∞	27.80 $\pm$15.69	**20.25** $\pm$4.81	**24.03**
SAM3D	2.83 $\pm$0.25	3.26 $\pm$0.24	2.12 $\pm$0.27	5.41 $\pm$0.22	3.60	263.51 $\pm$175.04	31.44 $\pm$21.84	28.31 $\pm$11.96	29.88
MaskSAM	1.86 $\pm$0.25	2.08 $\pm$0.20	2.12 $\pm$0.49	7.33 $\pm$0.30	3.84	26.18 $\pm$23.22	30.12 $\pm$16.51	22.92 $\pm$5.83	26.52
VFMs knowledge distillation methods									
KnowSAM	5.06 $\pm$0.47	4.95 $\pm$0.67	∞	10.48 $\pm$1.00	∞	∞	38.88 $\pm$14.58	33.06 $\pm$5.18	35.97
ESP-MedSAM	2.03 $\pm$0.25	2.36 $\pm$0.21	2.24 $\pm$0.43	7.66 $\pm$0.42	4.09	27.15 $\pm$45.21	37.84 $\pm$12.81	33.67 $\pm$4.65	35.76
IAD	7.17 $\pm$2.42	6.10 $\pm$0.64	8.97 $\pm$3.72	14.89 $\pm$3.19	9.99	161.05 $\pm$77.25	57.14 $\pm$11.87	71.02 $\pm$4.57	64.08
Embed-MedSAM	1.56 $\pm$0.25	1.79 $\pm$0.15	1.78 $\pm$0.35	1.91 $\pm$0.33	1.83	6.57 $\pm$7.30	25.16 $\pm$15.01	25.59 $\pm$6.43	25.37
BiLD (ours)	**1.04** $\pm$0.13	**1.13** $\pm$0.10	**1.10** $\pm$0.10	**1.07** $\pm$0.07	**1.10**	**1.24** $\pm$0.78	**10.94** $\pm$7.24	**5.83** $\pm$3.64	**8.39**

A.4 Case-Level Performance Distributions

Figure 12 shows Dice distributions for individual anatomical structures on three external brain datasets (OASIS, ADNI, and ABIDE) and two external cardiac datasets (MM-WHS and CAT08). These distributions complement the aggregate comparison in the main paper.

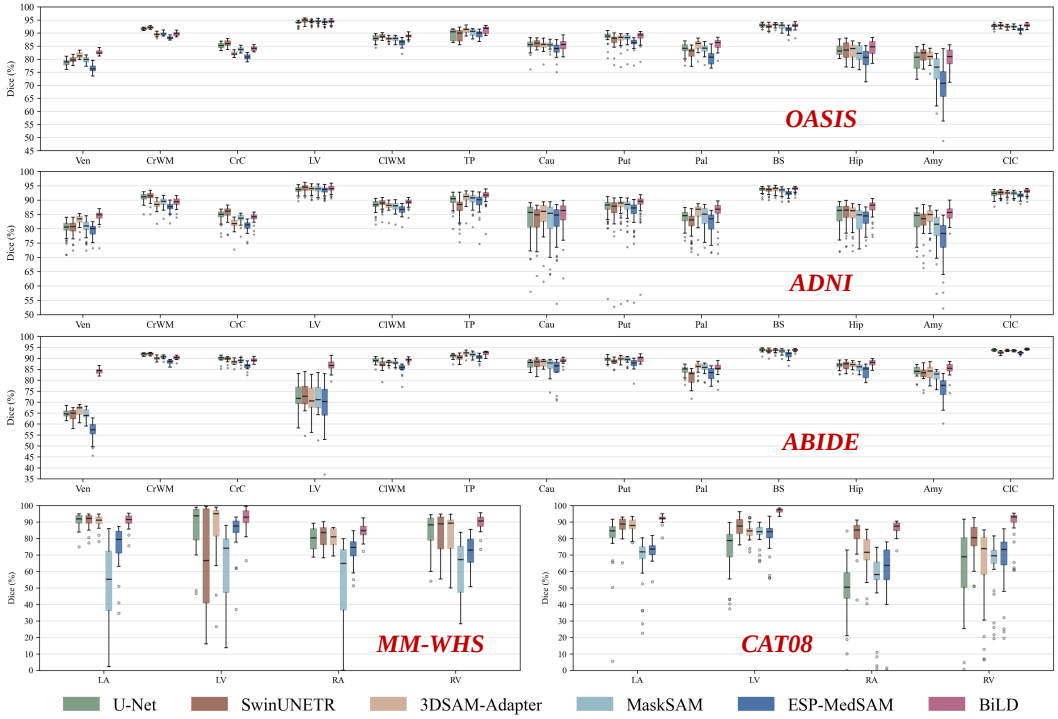


Figure 12: **Structure-wise Dice distributions on out-of-domain datasets.** The top three panels show OASIS, ADNI, and ABIDE; the bottom panels show MM-WHS and CAT08. Colors identify U-Net, SwinUNETR, 3DSAM-Adapter, MaskSAM, ESP-MedSAM, and BiLD. Higher Dice indicates better segmentation.

A.5 Qualitative Analysis of Teacher Configurations

Figure 13 complements the teacher configuration ablation in Table 2 of the main paper with representative cardiac segmentation examples. The visualization includes full teacher replacement with 3DSAM-Adapter and encoder replacement with DINOv2, MedSAM, BiomedParse, and MARS using a shared 3D decoder. TotalSeg is the source domain, while MM-WHS and CAT08 are external evaluation datasets. CAT08 is labeled Cardia32 in the supplied visualization.

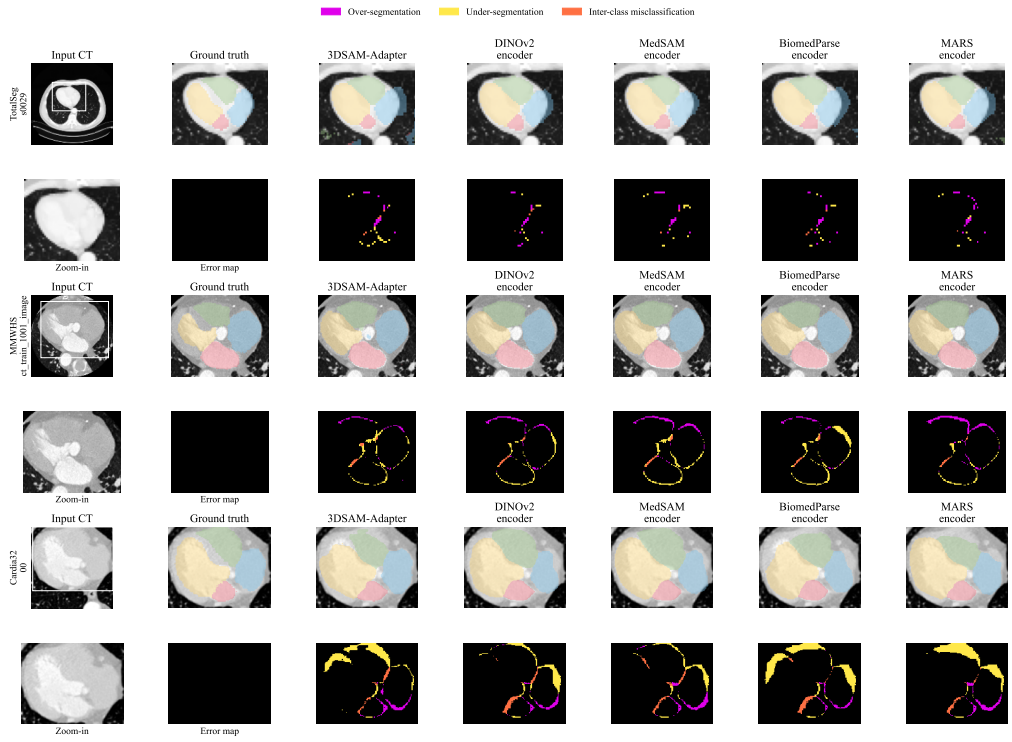


Figure 13: **Qualitative comparison of selected teacher configurations.** From top to bottom, the three case groups correspond to TotalSeg, MM-WHS, and CAT08 (labeled Cardia32 in the figure). Each group presents an input CT image, ground-truth labels, and segmentation overlays, with an enlarged input region and error maps below. The five prediction columns correspond to student results under the teacher configurations listed above each column. Magenta indicates over-segmentation, yellow indicates under-segmentation, and orange indicates inter-class misclassification. The figure covers a subset of the configurations in Table 2 of the main paper.

The displayed TotalSeg example has relatively sparse error regions, whereas the two external examples show more extensive discrepancies along anatomical boundaries. In the CAT08 example, the extent of the yellow under-segmentation regions varies across teacher configurations, while magenta and orange regions reveal additional boundary and inter-class errors. These spatial differences provide information that is not captured by a single average Dice score.

The quantitative results in Table 2 provide the dataset-level context: among the encoder replacements, MedSAM achieves the highest mean out-of-domain Dice (91.52%), followed by DINOv2 (91.16%) and MARS (90.96%). Together with the visual examples, these results indicate that teacher choice affects the transferred segmentation behavior and that BiLD can operate with multiple teacher architectures. The selected slices illustrate local error patterns rather than establish a ranking across the complete test sets.

References

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, Juan Eugenio Iglesias, and ADNI. SynthSeg: Segmentation of brain MRI scans

of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, 2023. doi: 10.1016/j.media.2023.102789. URL https://doi.org/10.1016/j.media.2023.102789.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pp. 424–432. Springer, 2016.

Xin Fan, Xiaolin Wang, Jiaxin Gao, Jia Wang, Zhongxuan Luo, and Risheng Liu. Bi-level learning of task-specific decoders for joint registration and one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11726–11735, 2024.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, 2024.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Kaiwen Huang, Tao Zhou, Huazhu Fu, Yizhe Zhang, Yi Zhou, Chen Gong, and Dong Liang. Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A. Landman. 3d UX-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=wsZsjOSytRA.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1540–1552. PMLR, 2020. URL https://proceedings.mlr.press/v108/lorraine20a.html.

Chonghua Lv, Dong Zhao, Shuang Wang, Dou Quan, Ning Huyan, Nicu Sebe, and Zhun Zhong. Generalizable knowledge distillation from vision foundation models for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.02554.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2023. doi: 10.1109/TMI.2022.3224067. URL https://doi.org/10.1109/TMI.2022.3224067.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11557–11568, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Pham_Meta_Pseudo_Labels_CVPR_2021_paper.html.

Rushi Qiang, Ruiyi Zhang, and Pengtao Xie. BiLoRA: A bi-level optimization framework for overfitting-resilient low-rank adaptation of large pre-trained models. *arXiv preprint arXiv:2403.13037*, 2024. URL https://arxiv.org/abs/2403.13037.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Qing Xu, Jiaxuan Li, Xiangjian He, Ziyu Liu, Zhen Chen, Wenting Duan, Chenxin Li, Maggie M He, Fiseha B Tesema, Wooi P Cheah, et al. Esp-medsam: Efficient self-prompting sam for universal domain-generalized medical image segmentation. *arXiv preprint arXiv:2407.14153*, 2024.