

---

# Appendix: Segment Anything Model Meets Semi-supervised Medical Image Segmentation: A Novel Perspective

---

Haifeng Zhao<sup>1,2</sup>, Haiyang Li<sup>1,2</sup>, Lei-Lei Ma<sup>1,2\*</sup>, Dengdi Sun<sup>2,3\*</sup>

<sup>1</sup>Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,  
School of Computer Science and Technology, Anhui University, China

<sup>2</sup>Key Lab Intelligent Comp & Signal Proc, Minist Educ, Anhui University, China

<sup>3</sup>School of Artificial Intelligence, Anhui University, China

In this appendix, Section A first presents the details of the datasets used in the experiments and the corresponding evaluation metrics. Then, in Section B, we show more details of the implementation of our method, including training and inference. In Section C, we describe the additional experimental analysis of our method, including the results of fine-tuning the teacher model, comparative and ablation studies on the PROMISE12 dataset [1], complexity comparisons between our teacher and student models, and comparisons of SAM and MedSAM as the teacher model. In addition, we provide in Section D a theoretical analysis of the proposed two-stage semi-supervised distillation framework, offering formal insights into how unlabeled data and knowledge transfer jointly influence generalization performance. Finally, in Section E, we provide a discussion covering the limitations and potential impacts of this work.

## A Datasets Details and Evaluation Metrics

**LA Dataset.** The left atrial (LA) dataset<sup>1</sup>, sourced from the 2018 Atrial Segmentation Challenge [2], comprises 100 3D gadolinium-enhanced magnetic resonance image scans, with 80 scans allocated for training and 20 for testing. Following previous work, four metrics are utilized: Dice Similarity Coefficient (DSC), Jaccard, 95% Hausdorff Distance (95HD), and Average Surface Distance (ASD).

**Brats-2019 Dataset.** The Brats-2019 dataset<sup>2</sup> [3] includes preoperative MRI scans from 335 glioma patients (259 high-grade and 76 low-grade) collected from multiple medical centers. Each scan consists of four modalities: T1, T1Gd, T2, and T2-FLAIR, along with pixel-wise annotations. In this work, we use only the T2-FLAIR images and split the dataset into 250 training cases, 25 validation cases, and 60 test cases. Evaluation metrics include DSC, Jaccard, 95HD, and ASD.

**PROMISE12 Dataset.** The PROMISE12 dataset<sup>3</sup>, introduced in the MICCAI 2012 Prostate Segmentation Challenge [1], contains MRI scans from 50 patients with diverse medical conditions across different locations. The dataset is divided into 35 cases for training, 5 for validation, and 10 for testing. As in prior research, DSC and ASD are chosen as evaluation metrics.

**Synapse Multi-Organ CT Dataset.** The Synapse dataset<sup>4</sup>, derived from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge [4], consists of 30 cases of abdominal CT scans. It is split into 18 cases for training and 12 cases for testing, with eight abdominal organs evaluated (aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas and stomach) using DSC and 95HD as metrics.

---

<sup>1</sup><https://www.cardiacatlas.org/atriaseg2018-challenge/>

<sup>2</sup><https://www.med.upenn.edu/cbica/brats-2019/>

<sup>3</sup><https://promise12.grand-challenge.org/Home/>

<sup>4</sup><https://www.synapse.org/>

## B Implementation Details

For the foundational teacher SAM, we conduct all experiments based on the “ViT-B” version<sup>5</sup>, while for the efficient student SAM, we replace the original image encoder with TinyViT-5M<sup>6</sup>. During fine-tuning, we apply LoRA [5] to the teacher SAM using the same configuration as SAMed [6], with the rank set to 4. The training loss combines Dice loss and Cross-Entropy loss, weighted at 0.8 and 0.2, respectively. We use the AdamW optimizer with  $\beta_1$  and  $\beta_2$  set to 0.9 and 0.999, and a weight decay of 0.1. The initial learning rates are set to 0.002 for the teacher SAM and 0.0005 for the student SAM, with a batch size of 8. In the semi-supervised learning phase, we employ the SGD optimizer with a weight decay of 0.0001 and an initial learning rate of 0.0023. The batch size is set to 16, comprising 8 labeled and 8 unlabeled data samples. The knowledge distillation loss assigns weight ratios of 1/3 and 2/3 to the embedding loss and the output loss, respectively. During training, all 3D scans are converted into 2D slices and upsampled to a resolution of  $512 \times 512$  for input. To ensure a fair comparison, we use data preprocessing and augmentation strategies consistent with previous studies. After training, we perform inference using only the efficient student SAM backbone, without requiring any prompts. All of our experiments are implemented in PyTorch with fixed random seeds, using a single NVIDIA RTX 3090 GPU in most cases, while two GPUs are used only for experiments on the Synapse dataset.

## C More Experiments

**Results on Fine-tuned Teacher Model.** In semi-supervised learning, the fine-tuned teacher model plays a crucial role in optimizing the student model. It not only provides effective supervisory guidance but also transfers valuable visual priors, thereby enhancing the student model’s performance. As shown in Table 1, we conduct experiments across multiple datasets to compare the original SAM with its fine-tuned counterpart. Results demonstrate that the fine-tuned SAM consistently outperforms the original model across all datasets. These findings confirm the effectiveness of our fine-tuning strategy in improving the teacher model’s capability and establish the fine-tuned SAM as a reliable teacher in the semi-supervised learning stage.

Table 1: Comparison between original SAM and fine-tuned SAM using DSC as the evaluation metric. The latter was fine-tuned with 10% labeled data.

Model	Dataset			
	LA	Brats-2019	PROMISE12	Synapse
Original SAM	17.59	6.47	2.53	3.48
Fine-tuned SAM	<b>88.54</b>	<b>84.04</b>	<b>85.39</b>	<b>75.99</b>

Table 2: Comparisons with SOTA semi-supervised segmentation methods on the PROMISE12 dataset.

Method	Scans used		Metrics	
	Labeled	Unlabeled	DSC↑	ASD↓
CCT [7]	7(20%)	28(80%)	71.43	16.61
URPC [8]			63.23	4.33
SS-Net [9]			62.31	4.36
SLC-Net [10]			68.31	4.69
SCP-Net [11]			77.06	3.52
BCP [12]			81.78	3.99
ABD [13]			82.06	1.33
Ours			<b>86.44</b>	<b>0.90</b>
ABD [13]	3(10%)	32(90%)	81.81	1.46
Ours			<b>83.41</b>	<b>1.15</b>

<sup>5</sup>[https://dl.fbaipublicfiles.com/segment\\_anything/sam\\_vit\\_b\\_01ec64.pth](https://dl.fbaipublicfiles.com/segment_anything/sam_vit_b_01ec64.pth)

<sup>6</sup><https://github.com/microsoft/Cream/tree/main/TinyViT>

**Results on PROMISE12 Dataset.** We perform experiments on the PROMISE12 dataset with 20% labeled ratio against CCT [7], URPC [8], SS-Net [9], SLC-Net [10], SCP-Net [11], BCP [12], and ABD [13], as well as ABD with 10% labeled ratio. As illustrated in Table 2, with 10% labeled data, our method significantly surpasses the second-best method (ABD) by 1.60% in DSC and 0.31 in ASD, demonstrating its superior segmentation capabilities. When using 20% labeled data, this performance gap increases further to 4.38% in DSC and 0.43 in ASD, highlighting the effectiveness of the proposed method under limited data conditions.

**Ablation Studies on PROMISE12 Dataset.** We also conduct some ablation experiments on the PROMISE12 dataset. Table 3 shows the detailed results of combining different knowledge distillation losses. Notably, using embedding loss and logit loss together yields the best performance, clearly demonstrating their complementary effects in improving the segmentation results. Table 4 shows the effectiveness of the proposed dynamic loss weighting (DLW) strategy. Applying DLW not only significantly improves DSC but also maintains stable ASD values, highlighting its role in effectively balancing the guidance from the teacher model during training.

Table 3: Ablation study on combinations of knowledge distillation losses on PROMISE12.

Embedding loss	Logit loss	Scans used		Metrics	
		Labeled	Unlabeled	DSC↑	ASD↓
✓				75.57	2.56
	✓	3(10%)	32(90%)	83.04	<b>1.10</b>
✓	✓			<b>83.41</b>	1.15
✓				83.06	1.12
	✓	7(20%)	28(80%)	85.64	1.02
✓	✓			<b>86.44</b>	<b>0.90</b>

Table 4: Effectiveness of the proposed DLW strategy on PROMISE12.

Method	Scans used		Metrics	
	Labeled	Unlabeled	DSC↑	ASD↓
w/o DLW	3(10%)	32(90%)	82.64	<b>1.13</b>
w/ DLW			<b>83.41</b>	1.15
w/o DLW	7(20%)	28(80%)	85.30	1.08
w/ DLW			<b>86.44</b>	<b>0.90</b>

**Complexity Analysis.** To demonstrate the deployability of the proposed method, we compare the model complexity of the foundational teacher SAM and the efficient student SAM backbone in our method on the PROMISE12 dataset. As shown in Table 5, the student model used for inference significantly reduces the number of parameters and Multiply-Accumulate Operations (MACs) for a single forward pass. By effectively leveraging valuable knowledge distilled from the high-complexity teacher on unlabeled data during semi-supervised learning, the student achieves improved performance while maintaining high efficiency. These results underscore the potential of the proposed method for practical clinical application, particularly in scenarios with limited computational resources.

Table 5: Complexity comparisons between our teacher SAM and student SAM.

Model	Complexity	
	Params (M)	MACs (G)
Foundational teacher SAM	105.05	103.40
Efficient student SAM backbone	<b>9.51</b>	<b>10.33</b>

**SAM vs. MedSAM.** We begin with the original SAM to validate the generality and transferability of our framework, along with the effectiveness of the proposed learning strategy. For comprehensive comparison, we also include experiments initializing with MedSAM [14] as the teacher model. As

shown in Table 6, we compare the zero-shot performance of SAM against MedSAM, where MedSAM is provided with bounding boxes derived from ground truth as prompts. The results confirm the strong generalizability of MedSAM in medical image segmentation. However, after fine-tuning with the same LoRA strategy, the performance of SAM even surpasses that of MedSAM. We attribute this phenomenon to the following reasons: although MedSAM undergoes domain-specific pre-training on medical images, its pre-training dataset size (approximately 1 million images) is considerably smaller than that of SAM, which is pre-trained on a large-scale natural image dataset (approximately 11 million images and 1 billion masks). During the subsequent task-specific fine-tuning stage with limited annotated data, the prior benefits brought by this “small-scale domain pre-training” are likely to be quickly offset by task-driven adaptation, and may even weaken the model’s inherent general representation capability [15]. Consequently, employing fine-tuned SAM as the teacher model yields superior performance when guiding the student model in semi-supervised learning.

Table 6: Comparison between SAM and MedSAM as the teacher model on the PROMISE12 dataset with 20% labeled data.

Stage	Model	Metrics	
		DSC↑	ASD↓
Zero-shot	MedSAM	<b>66.56</b>	<b>10.04</b>
	SAM (Ours)	2.53	94.65
After fine-tuning	MedSAM	81.26	3.79
	SAM (Ours)	<b>86.00</b>	<b>1.44</b>
Final results	MedSAM	85.13	1.56
	SAM (Ours)	<b>86.44</b>	<b>0.90</b>

## D Theoretical Analysis

This section provides a theoretical foundation for the proposed two-stage semi-supervised distillation framework. Our analysis follows the line of work that interprets knowledge distillation as a form of regularization or information transfer [16, 17], and builds upon uniform convergence tools from statistical learning theory to characterize the generalization behavior of the student model. Specifically, we show how incorporating the distillation term as a regularizer enables the student to leverage unlabeled data more effectively, thereby improving its generalization performance. The derived excess-risk bounds further clarify the conditions under which the student can achieve a lower expected risk than its teacher.

### D.1 Preliminaries

#### D.1.1 Notation and Definition

Let  $\mathcal{D}$  denote the underlying data distribution on  $\mathcal{X} \times \mathcal{Y}$ . We are given  $n_\ell$  (or is  $N$ ) labeled samples  $\mathcal{D}_\ell = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_\ell}$  and  $n_u$  (or is  $M$ ) unlabeled samples  $\mathcal{D}_u = \{\mathbf{x}_j\}_{j=1}^{n_u}$ , both drawn i.i.d. from  $\mathcal{D}$ . The teacher and student models are denoted by  $f_T \in \mathcal{F}_T$  and  $f_S \in \mathcal{F}_S$ , respectively. Each model induces a predictive distribution over labels:  $p_T(\mathbf{y}|\mathbf{x}) = f_T(\mathbf{x})$ ,  $p_S(\mathbf{y}|\mathbf{x}) = f_S(\mathbf{x})$ .

**Definition D.1** (Supervised Risk). *For a bounded loss function  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ , the expected risk and empirical risk of the student model  $f_S$  are defined as*

- *expected risk:*  $R(f_S) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}}[\ell(\mathbf{y}, p_S(\cdot|\mathbf{x}))]$
- *empirical risk:*  $\hat{R}_\ell(f_S) = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \ell(\mathbf{y}_i, p_S(\cdot|\mathbf{x}_i))$

**Definition D.2** (Knowledge Distillation Discrepancy). *For unlabeled data, the discrepancy between the teacher and student predictive distributions is measured by the Kullback–Leibler divergence:*

$$g_S(\mathbf{x}) = D_{\text{KL}}(p_T(\cdot|\mathbf{x}) \| p_S(\cdot|\mathbf{x})), \quad 0 \leq g_S(\mathbf{x}) \leq B,$$

where  $B > 0$  is a constant bounding the per-sample KL divergence. The expected and empirical distillation losses are given by

- *expected distillation loss*:  $\mathcal{L}_{\text{KD}}(f_S) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_u}[g_S(\mathbf{x})]$
- *empirical distillation loss*:  $\hat{\mathcal{L}}_{\text{KD}}(f_S) = \frac{1}{n_u} \sum_{j=1}^{n_u} g_S(\mathbf{x}_j)$

Combining the supervised (Definition D.1) and distillation (Definition D.2) objectives, the student model is trained by minimizing the composite empirical loss

$$\hat{J}(f_S) = \hat{R}_\ell(f_S) + \lambda \hat{\mathcal{L}}_{\text{KD}}(f_S), \quad (1)$$

where  $\lambda > 0$  controls the strength of knowledge distillation.

**Definition D.3** (Complexity Measures). *Let  $\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell)$  and  $\mathfrak{R}_{n_u}(\mathcal{G})$  denote the Rademacher complexities of the supervised and distillation loss classes:*

$$\mathcal{L}_\ell = \{(\mathbf{x}, \mathbf{y}) \mapsto \ell(\mathbf{y}, p_S(\cdot|\mathbf{x})) : f_S \in \mathcal{F}_S\}, \quad \mathcal{G} = \{\mathbf{x} \mapsto g_S(\mathbf{x}) : f_S \in \mathcal{F}_S\}.$$

**Definition D.4** (Approximation Gap). *The approximation gap between the student and the teacher is defined as*

$$\Delta_{\text{approx}} := \inf_{f_S \in \mathcal{F}_S} R(f_S) - R(f_T),$$

which quantifies the capacity difference between the student and teacher function classes.

### D.1.2 Some Facts

We recall a few standard facts (see, e.g., [18]):

**Fact D.5** (Rademacher complexity). *For a function class  $\mathcal{F}$  and samples  $\{\mathbf{z}_i\}_{i=1}^n$ , the empirical Rademacher complexity is*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\mathbf{z}_i) \right],$$

where  $\sigma_i$  are independent Rademacher variables with  $\mathbb{P}(\sigma_i = \pm 1) = 1/2$ . The population Rademacher complexity is defined as

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_Z [\hat{\mathfrak{R}}_n(\mathcal{F})].$$

**Fact D.6** (Symmetrization). *Let  $\mathcal{L} = \{\ell(f, \cdot) : f \in \mathcal{F}\}$  be a class of loss functions bounded in  $[0, 1]$ . Then*

$$\mathbb{E}_Z \left[ \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f)) \right] \leq 2 \mathfrak{R}_n(\mathcal{L}),$$

where  $R(f) = \mathbb{E}[\ell(f, Z)]$  and  $\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, \mathbf{z}_i)$ .

**Fact D.7** (Contraction). *Let  $\mathcal{F}$  be a class of functions and let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be an  $L_\phi$ -Lipschitz function satisfying  $\phi(0) = 0$ . Then the following inequality holds:*

$$\mathfrak{R}_n(\phi \circ \mathcal{F}) \leq L_\phi \mathfrak{R}_n(\mathcal{F}),$$

where  $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$ .

## D.2 Concentration and Uniform Convergence Lemmas

To analyze the generalization behavior of the student model, we first recall one classical concentration inequality that will be used to control deviations between empirical and expected quantities. These probabilistic tools form the foundation for the uniform convergence results presented later in this subsection.

**Lemma D.8** (McDiarmid's Inequality [19]). *If the following conditions hold:*

- $X_1, \dots, X_n$  are independent random variables taking values in some set  $\mathcal{X}$ ,
- $F : \mathcal{X}^n \rightarrow \mathbb{R}$  is a measurable function,

- $F$  satisfies the bounded differences condition: for each  $i \in \{1, \dots, n\}$  and all  $x_1, \dots, x_n, \mathbf{x}'_i \in \mathcal{X}$ ,

$$|F(x_1, \dots, \mathbf{x}_i, \dots, x_n) - F(x_1, \dots, \mathbf{x}'_i, \dots, x_n)| \leq c_i,$$

then, for all  $t > 0$ ,

$$\mathbb{P}(F(X_1, \dots, X_n) - \mathbb{E}[F(X_1, \dots, X_n)] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

The above inequalities will be repeatedly used to bound the fluctuations of empirical risks around their expectations. We now apply them to the supervised and distillation objectives to obtain uniform convergence results.

**Lemma D.9** (Supervised uniform convergence). *If the following conditions hold:*

- $\mathcal{L}_\ell = \{(\mathbf{x}, \mathbf{y}) \mapsto \ell(\mathbf{y}, f(\mathbf{x})) : f \in \mathcal{F}_S\}$  is the induced loss class,
- $\delta \in (0, 1)$  is a confidence parameter,

then, with probability at least  $1 - \delta$  (over the sampling of  $\mathcal{D}_\ell$ ), the following holds simultaneously for all  $f \in \mathcal{F}_S$ :

$$R(f) \leq \hat{R}_\ell(f) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + \sqrt{\frac{\ln(2/\delta)}{2n_\ell}}. \quad (2)$$

*Proof sketch.* This follows directly from the Rademacher complexity bound for bounded losses (e.g. [18]). Since  $\ell \in [0, 1]$ , symmetrization and concentration yield  $\mathbb{E}[\sup_f (R(f) - \hat{R}_\ell(f))] \leq 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell)$ , and applying McDiarmid's inequality gives the stated high-probability bound.  $\square$

**Lemma D.10** (Distillation uniform convergence). *If the following conditions hold:*

- $\mathcal{G} = \{\mathbf{x} \mapsto g_f(\mathbf{x}) : f \in \mathcal{F}_S\}$  is the induced function class,
- $g_f(\mathbf{x}) \in [0, B]$  for all  $f \in \mathcal{F}_S$  and all  $\mathbf{x}$ ,
- $\delta \in (0, 1)$  is a confidence parameter,

then, with probability at least  $1 - \delta$  (over the draw of  $\mathcal{D}_u$ ), the following holds simultaneously for all  $f \in \mathcal{F}_S$ :

$$\mathcal{L}_{\text{KD}}(f) \leq \hat{\mathcal{L}}_{\text{KD}}(f) + 2\mathfrak{R}_{n_u}(\mathcal{G}) + B\sqrt{\frac{\ln(2/\delta)}{2n_u}}. \quad (3)$$

*Proof sketch.* The proof mirrors Lemma D.9, after normalizing by  $B$  and applying standard symmetrization and Hoeffding-type concentration for bounded functions.  $\square$

Together, Lemma D.9 and Lemma D.10 guarantee that both the supervised and distillation losses enjoy uniform convergence at rates determined by their respective sample sizes  $(n_\ell, n_u)$  and function class complexities. These results will be combined in the next subsection to derive the generalization bound for the overall objective (Eq. (1)).

### D.3 A Combined Bound

Having established uniform convergence results for both the supervised and distillation components, we now combine them to obtain a generalization guarantee for the overall composite objective  $\hat{J}(f)$  introduced in Eq. (1). This bound connects the empirical training loss with the true population risk, revealing how the two learning sources (labeled and unlabeled data) jointly influence generalization.

**Lemma D.11** (Composite objective generalization). *If the following conditions hold:*

- $\lambda > 0$  is a fixed regularization parameter,
- $\hat{J}(f) = \hat{R}_\ell(f) + \lambda \hat{\mathcal{L}}_{\text{KD}}(f)$  is the composite empirical objective,

- $\delta \in (0, 1)$  is a confidence parameter,

then, with probability at least  $1 - \delta$  (over the sampling of  $\mathcal{D}_\ell$  and  $\mathcal{D}_u$ ), the following inequality holds simultaneously for all  $f \in \mathcal{F}_S$ :

$$R(f) \leq \hat{J}(f) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2n_\ell}} + \lambda B \sqrt{\frac{\ln(2/\delta)}{2n_u}} - \lambda\mathcal{L}_{\text{KD}}(f). \quad (4)$$

*Proof sketch.* Apply Lemma D.9 and Lemma D.10 with confidence parameters  $\delta/2$ , then take a union bound. Rearranging  $\hat{\mathcal{L}}_{\text{KD}}(f) \geq \mathcal{L}_{\text{KD}}(f) - 2\mathfrak{R}_{n_u}(\mathcal{G}) - B\sqrt{\ln(4/\delta)/(2n_u)}$  and substituting  $\hat{R}_\ell(f) = \hat{J}(f) - \lambda\hat{\mathcal{L}}_{\text{KD}}(f)$  yields Eq. (4).  $\square$

The bound in Lemma D.11 provides a unified control of the population risk  $R(f)$  in terms of the empirical objective  $\hat{J}(f)$ . The additional terms capture the contribution of supervised and unlabeled samples via their respective complexities  $\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell)$  and  $\mathfrak{R}_{n_u}(\mathcal{G})$ . Notably, the negative term  $-\lambda\mathcal{L}_{\text{KD}}(f)$  highlights the beneficial effect of distillation: a smaller teacher–student discrepancy  $\mathcal{L}_{\text{KD}}(f)$  leads to a tighter generalization bound. This combined result serves as the foundation for the main theorem in the next subsection. This combined analysis connects our framework to the generalization theory of semi-supervised learning [20, 21], highlighting how unlabeled data can tighten risk bounds through the regularizing effect of knowledge transfer.

#### D.4 Main Result: Student vs. Teacher

Having established the uniform convergence and composite bounds, we now present the main theoretical result that compares the generalization performance of the student and teacher models. The theorem below quantifies how the student’s expected risk deviates from that of the teacher, highlighting the roles of approximation capacity, estimation complexity, and knowledge distillation.

**Theorem D.12** (Excess risk bound for semi-supervised distillation). *If the following conditions hold:*

- The loss function is bounded:  $\ell(\mathbf{y}, \hat{\mathbf{y}}) \in [0, 1]$ ,
- The distillation function is bounded:  $g_f(\mathbf{x}) \in [0, B]$  for all  $f \in \mathcal{F}_S$  and all  $\mathbf{x}$ ,
- The student  $f_S$  satisfies the approximate empirical minimization condition

$$\hat{J}(f_S) \leq \inf_{f \in \mathcal{F}_S} \hat{J}(f) + \varepsilon_{\text{opt}}$$

for some  $\varepsilon_{\text{opt}} \geq 0$ ,

- $\delta \in (0, 1)$  is a confidence parameter,

then, with probability at least  $1 - \delta$ , the following inequality holds:

$$\begin{aligned} R(f_S) &\leq R(f_T) + \Delta_{\text{approx}} + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) \\ &\quad + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda \frac{B \ln(4/\delta)}{n_u} + \varepsilon_{\text{opt}} - \lambda\mathcal{L}_{\text{KD}}(f_S). \end{aligned} \quad (5)$$

*Proof sketch.* Apply Lemma D.11 to  $f = f_S$  and use the near-optimality of  $f_S$ . Note that  $\inf_{f \in \mathcal{F}_S} \hat{J}(f) \leq \hat{J}(f_T) = \hat{R}_\ell(f_T)$  as  $\hat{\mathcal{L}}_{\text{KD}}(f_T) = 0$ . Use Lemma D.9 to bound  $\hat{R}_\ell(f_T)$  in terms of  $R(f_T)$ , substitute, and introduce  $\Delta_{\text{approx}} = \inf_{f \in \mathcal{F}_S} R(f) - R(f_T)$  to obtain Eq. (5).  $\square$

Theorem D.12 characterizes how the student’s generalization depends on approximation, estimation, and distillation terms. A direct consequence of this bound is a simple sufficient condition under which the student achieves a strictly lower expected risk than the teacher.

**Corollary D.13** (Sufficient condition for  $R(f_S) < R(f_T)$ ). *If the following conditions hold:*

- The assumptions of Theorem D.12 are satisfied,
- The inequality

$$\lambda\mathcal{L}_{\text{KD}}(f_S) > \Delta_{\text{approx}} + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda \frac{B \ln(4/\delta)}{n_u} + \varepsilon_{\text{opt}}$$

holds,

then, with probability at least  $1 - \delta$ , we have

$$R(f_S) < R(f_T).$$

*Proof sketch.* Immediate by rearranging Eq. (5).  $\square$

**Discussion.** The bound in Theorem D.12 decomposes the student's excess risk into four intuitive components: (i) an approximation term  $\Delta_{\text{approx}}$  reflecting the model capacity gap; (ii) estimation terms governed by the Rademacher complexities of the supervised and distillation loss classes; (iii) an optimization tolerance  $\varepsilon_{\text{opt}}$ ; and (iv) a beneficial term  $-\lambda\mathcal{L}_{\text{KD}}(f_S)$  due to distillation. When the effective knowledge transfer  $\lambda\mathcal{L}_{\text{KD}}(f_S)$  dominates the approximation and estimation errors, the student outperforms its teacher in terms of expected risk with high probability, consistent with prior analyses of distillation as an implicit regularization process [16, 17]. When the weighted distillation gain  $\lambda\mathcal{L}_{\text{KD}}(f_S)$  dominates the approximation and estimation errors, the student can achieve a lower expected risk than the teacher with high probability. Similar interpretations have been discussed in the recent theoretical studies of knowledge distillation [22, 23]

## D.5 Full proofs and explicit complexity bounds

This subsection provides complete derivations and additional theoretical analysis for Lemma D.9 through Theorem D.12. For completeness, we detail the intermediate steps involved in the symmetrization, contraction, and concentration arguments, and we further instantiate the abstract Rademacher complexity terms with more interpretable upper bounds based on parameter norms and Lipschitz continuity assumptions.

### D.5.1 Detailed Proof of Lemma D.9 (Supervised Uniform Convergence)

*Proof of Lemma D.9.* We restate the setting: the loss  $\ell(\mathbf{y}, f(\mathbf{x}))$  is bounded in  $[0, 1]$ , and  $\mathcal{L} = \{(\mathbf{x}, \mathbf{y}) \mapsto \ell(\mathbf{y}, f(\mathbf{x})) : f \in \mathcal{F}_S\}$ . Let  $\mathcal{D}_\ell = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_\ell}$  be i.i.d. draws from  $\mathcal{D}$ . The goal is to upper bound the deviation between population and empirical risks:

$$Z(\mathcal{D}_\ell) := \sup_{f \in \mathcal{F}_S} (R(f) - \hat{R}_\ell(f)) = \sup_{f \in \mathcal{F}_S} \left( \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} [\ell(\mathbf{y}, f(\mathbf{x}))] - \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \ell(\mathbf{y}_i, f(\mathbf{x}_i)) \right).$$

**Setp 1: symmetrization.** To control the expectation of  $Z(\mathcal{D}_\ell)$ , we introduce an independent ghost sample  $\mathcal{D}'_\ell = \{(\mathbf{x}'_i, \mathbf{y}'_i)\}_{i=1}^{n_\ell}$  drawn i.i.d. from  $\mathcal{D}$ . Then

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_\ell} [Z(\mathcal{D}_\ell)] &= \mathbb{E}_{\mathcal{D}_\ell} \left[ \sup_f \left( \mathbb{E}_{\mathcal{D}} [\ell(\mathbf{y}, f(\mathbf{x}))] - \hat{R}_\ell(f) \right) \right] \\ &\leq \mathbb{E}_{\mathcal{D}_\ell, \mathcal{D}'_\ell} \left[ \sup_f \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} (\ell(\mathbf{y}'_i, f(\mathbf{x}'_i)) - \ell(\mathbf{y}_i, f(\mathbf{x}_i))) \right]. \end{aligned}$$

Introducing Rademacher variables  $\sigma_i \in \{\pm 1\}$  with  $\mathbb{E}[\sigma_i] = 0$  and applying the standard symmetrization trick (Fact D.6), we obtain

$$\mathbb{E}_{\mathcal{D}_\ell} [Z(\mathcal{D}_\ell)] \leq 2 \mathbb{E}_{\mathcal{D}_\ell, \sigma} \left[ \sup_{f \in \mathcal{F}_S} \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \sigma_i \ell(\mathbf{y}_i, f(\mathbf{x}_i)) \right] = 2 \mathfrak{R}_{n_\ell}(\mathcal{L}_\ell),$$

which bounds the expected deviation by twice the empirical Rademacher complexity.

**Setp 2: concentration via McDiarmid's inequality.** Next, we control the deviation of  $Z(\mathcal{D}_\ell)$  around its mean. Changing a single example  $(\mathbf{x}_i, \mathbf{y}_i)$  to  $(\mathbf{x}'_i, \mathbf{y}'_i)$  alters  $\hat{R}_\ell(f)$  by at most  $1/n_\ell$ , since  $\ell(\cdot, \cdot) \in [0, 1]$ . Hence, for all  $i$ , the bounded-difference constant is  $c_i = 1/n_\ell$ . Applying McDiarmid's inequality (Lemma D.8), for any  $t > 0$ ,

$$\Pr[Z(\mathcal{D}_\ell) - \mathbb{E}_{\mathcal{D}_\ell} [Z(\mathcal{D}_\ell)] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n_\ell} c_i^2}\right) = \exp(-2t^2 n_\ell).$$

Setting the right-hand side equal to  $\delta$  and solving for  $t$  gives  $t = \sqrt{\ln(1/\delta)/(2n_\ell)}$ .



**Setp 3: combining results.** With probability at least  $1 - \delta$ ,

$$Z(\mathcal{D}_\ell) \leq \mathbb{E}_{\mathcal{D}_\ell}[Z(\mathcal{D}_\ell)] + \sqrt{\frac{\ln(1/\delta)}{2n_\ell}} \leq 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + \sqrt{\frac{\ln(1/\delta)}{2n_\ell}}.$$

To obtain a two-sided deviation, we replace  $\delta$  by  $\delta/2$ , yielding the final bound:

$$\sup_{f \in \mathcal{F}_S} (R(f) - \hat{R}_\ell(f)) \leq 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + \sqrt{\frac{\ln(2/\delta)}{2n_\ell}},$$

which concludes the proof.  $\square$

### D.5.2 Detailed Proof of Lemma D.10 (KD Uniform Convergence)

*Proof of Lemma D.10.* Let  $\mathcal{D}_u = \{\mathbf{x}_j\}_{j=1}^{n_u}$  be i.i.d. draws from  $\mathcal{D}$ , and recall that  $g_f(\mathbf{x}) = D_{\text{KL}}(p_T(\cdot|\mathbf{x}) \| p_f(\cdot|\mathbf{x})) \in [0, B]$ . Define the deviation

$$Z_u(\mathcal{D}_u) := \sup_{f \in \mathcal{F}_S} (\mathcal{L}_{\text{KD}}(f) - \hat{\mathcal{L}}_{\text{KD}}(f)) = \sup_{f \in \mathcal{F}_S} \left( \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[g_f(\mathbf{x})] - \frac{1}{n_u} \sum_{j=1}^{n_u} g_f(\mathbf{x}_j) \right).$$

**Setp 1: symmetrization.** Introduce an independent ghost sample  $\mathcal{D}'_u = \{\mathbf{x}'_j\}_{j=1}^{n_u}$  drawn i.i.d. from  $\mathcal{D}$ . Then

$$\mathbb{E}_{\mathcal{D}_u}[Z_u(\mathcal{D}_u)] \leq \mathbb{E}_{\mathcal{D}_u, \mathcal{D}'_u} \left[ \sup_{f \in \mathcal{F}_S} \frac{1}{n_u} \sum_{j=1}^{n_u} (g_f(\mathbf{x}'_j) - g_f(\mathbf{x}_j)) \right].$$

Let  $\sigma_j \in \{\pm 1\}$  be independent Rademacher variables with  $\mathbb{E}[\sigma_j] = 0$ . Using the standard symmetrization trick (Fact D.6),

$$\mathbb{E}_{\mathcal{D}_u}[Z_u(\mathcal{D}_u)] \leq 2 \mathbb{E}_{\mathcal{D}_u, \sigma} \left[ \sup_{f \in \mathcal{F}_S} \frac{1}{n_u} \sum_{j=1}^{n_u} \sigma_j g_f(\mathbf{x}_j) \right] = 2\mathfrak{R}_{n_u}(\mathcal{G}),$$

where  $\mathcal{G} = \{\mathbf{x} \mapsto g_f(\mathbf{x}) : f \in \mathcal{F}_S\}$ .

**Setp 2: concentration via McDiarmid's inequality.** Changing one sample  $\mathbf{x}_j$  to  $\mathbf{x}'_j$  changes  $\hat{\mathcal{L}}_{\text{KD}}(f)$  by at most  $B/n_u$  for any  $f$ , since  $g_f(\mathbf{x}) \in [0, B]$ . Thus,  $Z_u(\mathcal{D}_u)$  satisfies the bounded-difference condition with constants  $c_j \leq B/n_u$ . Applying McDiarmid's inequality (Lemma D.8) yields, for any  $t > 0$ ,

$$\Pr[Z_u(\mathcal{D}_u) - \mathbb{E}_{\mathcal{D}_u}[Z_u(\mathcal{D}_u)] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^{n_u} c_j^2}\right) = \exp\left(-\frac{2t^2 n_u}{B^2}\right).$$

Setting the right-hand side to  $\delta$  gives  $t = B\sqrt{\ln(1/\delta)/(2n_u)}$ .

**Setp 3: combining results.** With probability at least  $1 - \delta$ ,

$$Z_u(\mathcal{D}_u) \leq \mathbb{E}_{\mathcal{D}_u}[Z_u(\mathcal{D}_u)] + B\sqrt{\frac{\ln(1/\delta)}{2n_u}} \leq 2\mathfrak{R}_{n_u}(\mathcal{G}) + B\sqrt{\frac{\ln(1/\delta)}{2n_u}}.$$

Finally, replacing  $\delta$  with  $\delta/2$  for a one-sided deviation bound, we obtain that, with probability at least  $1 - \delta$ ,

$$\mathcal{L}_{\text{KD}}(f) \leq \hat{\mathcal{L}}_{\text{KD}}(f) + 2\mathfrak{R}_{n_u}(\mathcal{G}) + B\sqrt{\frac{\ln(2/\delta)}{2n_u}},$$

which is exactly the claim in Eq. (3).  $\square$

### D.5.3 Detailed Proof of Lemma D.11 (Combined bound)

*Proof of Lemma D.11.* We prove the claimed inequality by combining the supervised and KD uniform-convergence bounds and taking a union bound over the two sampled sets.

Fix  $\delta \in (0, 1)$ . Apply Lemma D.9 to the labeled sample  $\mathcal{D}_\ell$  with confidence parameter  $\delta/2$ . By Lemma D.9, with probability at least  $1 - \delta/2$  (over  $\mathcal{D}_\ell$ ) we have for all  $f \in \mathcal{F}_S$ ,

$$R(f) \leq \hat{R}_\ell(f) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + \sqrt{\frac{\ln(2/(\delta/2))}{2n_\ell}} = \hat{R}_\ell(f) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}}. \quad (6)$$

Similarly, apply Lemma D.10 to the unlabeled sample  $\mathcal{D}_u$  with confidence parameter  $\delta/2$ . By Lemma D.10, with probability at least  $1 - \delta/2$  (over  $\mathcal{D}_u$ ) we have for all  $f \in \mathcal{F}_S$ ,

$$\mathcal{L}_{\text{KD}}(f) \leq \hat{\mathcal{L}}_{\text{KD}}(f) + 2\mathfrak{R}_{n_u}(\mathcal{G}) + B\sqrt{\frac{\ln(2/(\delta/2))}{2n_u}} = \hat{\mathcal{L}}_{\text{KD}}(f) + 2\mathfrak{R}_{n_u}(\mathcal{G}) + B\sqrt{\frac{\ln(4/\delta)}{2n_u}}. \quad (7)$$

By a union bound, both Eq. (6) and Eq. (7) hold simultaneously with probability at least  $1 - \delta$  (over the draw of  $\mathcal{D}_\ell$  and  $\mathcal{D}_u$ ).

From Eq. (7) we obtain a corresponding lower bound on the empirical KD term:

$$\hat{\mathcal{L}}_{\text{KD}}(f) \geq \mathcal{L}_{\text{KD}}(f) - 2\mathfrak{R}_{n_u}(\mathcal{G}) - B\sqrt{\frac{\ln(4/\delta)}{2n_u}}.$$

Substitute  $\hat{R}_\ell(f) = \hat{J}(f) - \lambda\hat{\mathcal{L}}_{\text{KD}}(f)$  into Eq. (6) and then replace  $\hat{\mathcal{L}}_{\text{KD}}(f)$  by the lower bound just obtained. Carrying out these substitutions yields, with probability at least  $1 - \delta$ ,

$$\begin{aligned} R(f) &\leq \hat{J}(f) - \lambda\hat{\mathcal{L}}_{\text{KD}}(f) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} \\ &\leq \hat{J}(f) - \lambda\left(\mathcal{L}_{\text{KD}}(f) - 2\mathfrak{R}_{n_u}(\mathcal{G}) - B\sqrt{\frac{\ln(4/\delta)}{2n_u}}\right) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} \\ &= \hat{J}(f) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda B\sqrt{\frac{\ln(4/\delta)}{2n_u}} - \lambda\mathcal{L}_{\text{KD}}(f). \end{aligned}$$

This is precisely the claim Eq. (4), concluding the proof.  $\square$

### D.5.4 Detailed Proof of Theorem D.12 (Excess risk bound for semi-supervised distillation)

*Proof of Theorem D.12.* We give a self-contained, step-by-step derivation that makes explicit the choice of confidence parameters and the algebraic substitutions.

**Step 1: setup and goal.** Recall the combined bound of Lemma D.11: for any  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$  (over the sampling of  $\mathcal{D}_\ell$  and  $\mathcal{D}_u$ ), every  $f \in \mathcal{F}_S$  satisfies

$$\begin{aligned} R(f) &\leq \hat{J}(f) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) \\ &\quad + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda B\sqrt{\frac{\ln(4/\delta)}{2n_u}} - \lambda\mathcal{L}_{\text{KD}}(f). \end{aligned} \quad (8)$$

We will apply Eq. (8) to  $f = f_S$ , then replace  $\hat{J}(f_S)$  by an (approximate) empirical optimum and relate that optimum to  $R(f_T)$  to produce the desired excess-risk bound.

**Step 2: apply Lemma D.11 to  $f_S$ .** Fix  $\delta \in (0, 1)$ . By Lemma D.11 (with this  $\delta$ ), with probability at least  $1 - \delta$ ,

$$\begin{aligned} R(f_S) &\leq \hat{J}(f_S) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) \\ &\quad + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda B\sqrt{\frac{\ln(4/\delta)}{2n_u}} - \lambda\mathcal{L}_{\text{KD}}(f_S). \end{aligned}$$

**Step 3: use the (approximate) empirical optimality of  $f_S$ .** By assumption,

$$\hat{J}(f_S) \leq \inf_{f \in \mathcal{F}_S} \hat{J}(f) + \varepsilon_{\text{opt}},$$

so with the same probability event we can substitute to obtain

$$\begin{aligned} R(f_S) &\leq \inf_{f \in \mathcal{F}_S} \hat{J}(f) + \varepsilon_{\text{opt}} + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) \\ &\quad + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda B \sqrt{\frac{\ln(4/\delta)}{2n_u}} - \lambda \mathcal{L}_{\text{KD}}(f_S). \end{aligned} \quad (9)$$

**Step 4: upper-bound  $\inf_f \hat{J}(f)$  by  $\hat{J}(f_T)$  and relate to  $R(f_T)$ .** Observe that for the teacher model  $f_T$  we have  $\hat{\mathcal{L}}_{\text{KD}}(f_T) = 0$ , since the distillation loss measures the discrepancy between the teacher and student predictions, and for  $f_T$  this discrepancy vanishes (i.e.,  $D_{\text{KL}}(p_T \| p_T) = 0$ ). Hence,  $\hat{J}(f_T) = \hat{R}_\ell(f_T)$ . Therefore

$$\inf_{f \in \mathcal{F}_S} \hat{J}(f) \leq \hat{J}(f_T) = \hat{R}_\ell(f_T).$$

Substitute this inequality into Eq. (9) to get

$$\begin{aligned} R(f_S) &\leq \hat{R}_\ell(f_T) + \varepsilon_{\text{opt}} + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) \\ &\quad + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda B \sqrt{\frac{\ln(4/\delta)}{2n_u}} - \lambda \mathcal{L}_{\text{KD}}(f_S). \end{aligned} \quad (10)$$

We now relate  $\hat{R}_\ell(f_T)$  to its population counterpart  $R(f_T)$ . Apply Lemma D.9 to the teacher function  $f_T$  with the same confidence bookkeeping used previously (this can be done so that all events hold simultaneously with probability at least  $1 - \delta$ ; one convenient approach is to use the same  $\delta/2$  split used in the proof of Lemma D.11, producing the  $\ln(4/\delta)$  terms). Lemma D.9 yields (with the appropriate confidence)

$$\hat{R}_\ell(f_T) \leq R(f_T) + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}}.$$

Substituting this into Eq. (10) gives

$$\begin{aligned} R(f_S) &\leq R(f_T) + \underbrace{\inf_{f \in \mathcal{F}_S} R(f) - R(f_T)}_{\Delta_{\text{approx}}} + 2\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) + 2\lambda\mathfrak{R}_{n_u}(\mathcal{G}) \\ &\quad + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda B \sqrt{\frac{\ln(4/\delta)}{2n_u}} + \varepsilon_{\text{opt}} - \lambda \mathcal{L}_{\text{KD}}(f_S). \end{aligned}$$

Here we used the identity  $\inf_{f \in \mathcal{F}_S} R(f) = R(f_T) + \Delta_{\text{approx}}$ , which defines  $\Delta_{\text{approx}}$ .

**Step 5: collect terms and conclude.** Collecting the displayed terms yields exactly the bound (5) stated in the theorem. All probabilistic statements above can be arranged to hold simultaneously by appropriately applying the union bound when invoking the lemmas; this bookkeeping is the place where the  $\ln(4/\delta)$  factors arise (as detailed in the proof of Lemma D.11).

This completes the proof of Theorem D.12.  $\square$

### D.5.5 Replacing abstract Rademacher terms with Lipschitz & parameter-norm bounds

The abstract Rademacher complexities  $\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell)$  and  $\mathfrak{R}_{n_u}(\mathcal{G})$  can be upper-bounded in terms of model smoothness and parameter norms. This subsection presents two practical scenarios and corresponding interpretable bounds, while emphasizing the dependence on the labeled and unlabeled sample sizes  $(n_\ell, n_u)$ .

**Case A: Linear predictors with bounded feature norm.** Assume the student model is linear in parameters,  $f_\theta(\mathbf{x}) = \langle \theta, \phi(\mathbf{x}) \rangle$ , where  $\phi(\mathbf{x}) \in \mathbb{R}^d$ ,  $\|\theta\|_2 \leq S$  for all  $\theta \in \Theta$ , and  $\|\phi(\mathbf{x})\|_2 \leq F$  for all  $\mathbf{x}$ . Suppose the supervised loss  $\ell(\mathbf{y}, \hat{\mathbf{y}})$  is  $L_\ell$ -Lipschitz in its second argument, and the distillation function  $g_f(\mathbf{x})$  is  $L_g$ -Lipschitz in the scalar network output. Then, standard results (see [18]) give

$$\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) \leq L_\ell \mathfrak{R}_{n_\ell}(\{\mathbf{x} \mapsto \langle \theta, \phi(\mathbf{x}) \rangle : \|\theta\| \leq S\}) \leq L_\ell \frac{SF}{\sqrt{n_\ell}},$$

and, similarly for the distillation class,

$$\mathfrak{R}_{n_u}(\mathcal{G}) \leq L_g \frac{SF}{\sqrt{n_u}}.$$

Hence, one can set  $C_\ell = L_\ell SF$  and  $C_g = L_g SF$  in the unified substitution below.

**Case B: Deep networks controlled by spectral norms.** For an  $L$ -layer feedforward network with weight matrices  $W_1, \dots, W_L$  and 1-Lipschitz activations (e.g., ReLU), the overall Lipschitz constant satisfies  $L_f \leq \prod_{i=1}^L \|W_i\|_2$ . If  $\|x\| \leq X$  and each  $\|W_i\|_2 \leq s_i$ , then empirical-process arguments (e.g., 24, 25, 26) imply that, up to logarithmic factors,

$$\mathfrak{R}_n(\mathcal{F}) \lesssim \frac{X \prod_{i=1}^L s_i}{\sqrt{n}},$$

where  $n$  should be replaced by  $n_\ell$  or  $n_u$  depending on the corresponding loss term. Consequently,

$$\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) \lesssim L_\ell \frac{X \prod_{i=1}^L s_i}{\sqrt{n_\ell}}, \quad \mathfrak{R}_{n_u}(\mathcal{G}) \lesssim L_g \frac{X \prod_{i=1}^L s_i}{\sqrt{n_u}}.$$

These bounds highlight how spectral-norm control and input magnitude determine the effective complexity of the model.

**Unified substitution and explicit bound.** Let constants  $C_\ell, C_g > 0$  capture the dependence on parameter norms, input norms, network depth, and Lipschitz constants, such that

$$\mathfrak{R}_{n_\ell}(\mathcal{L}_\ell) \leq \frac{C_\ell}{\sqrt{n_\ell}}, \quad \mathfrak{R}_{n_u}(\mathcal{G}) \leq \frac{C_g}{\sqrt{n_u}}.$$

Substituting these estimates into the excess-risk bound (5) gives the explicit, sample-dependent inequality

$$\begin{aligned} R(f_S) &\leq R(f_T) + \Delta_{\text{approx}} + 2 \frac{C_\ell}{\sqrt{n_\ell}} + 2\lambda \frac{C_g}{\sqrt{n_u}} + \sqrt{\frac{\ln(4/\delta)}{2n_\ell}} + \lambda B \sqrt{\frac{\ln(4/\delta)}{2n_u}} \\ &\quad + \varepsilon_{\text{opt}} - \lambda \mathcal{L}_{\text{KD}}(f_S). \end{aligned} \tag{11}$$

**Remarks.**

- **Sample dependence.** The supervised term scales with  $n_\ell^{-1/2}$ , while the distillation term scales with  $n_u^{-1/2}$ , reflecting the separate contributions of labeled and unlabeled data to generalization.
- **Concentration term correction.** The KD-related concentration term must retain the square-root form  $\lambda B \sqrt{\ln(4/\delta)/(2n_u)}$ , consistent with the McDiarmid inequality, rather than a linear-inverse form.
- **Interpretability.** The constants  $C_\ell$  and  $C_g$  summarize the interaction between model capacity (e.g., spectral or Euclidean norms), input scale, and Lipschitz smoothness of the loss and KD terms. Larger norms or smaller sample sizes enlarge the bound, while stronger Lipschitz control and regularization improve it.

#### D.5.6 Practical notes on constants and empirical checks

- **Estimating  $C_\ell, C_g$ .** For linear models,  $C_\ell \approx L_\ell SF$  and  $C_g \approx L_g SF$  where  $S$  is the weight-norm bound and  $F$  the feature-norm bound. For deep neural networks,  $C_\ell$  can be upper bounded by  $L_\ell X \prod_i s_i$  (up to logarithmic factors).
- **Behavior as  $n_u$  grows.** The distillation estimation term scales like  $O(\lambda C_g / \sqrt{n_u})$  while the beneficial term is  $-\lambda \mathcal{L}_{\text{KD}}(f_S)$ . If increasing  $n_u$  reduces  $\mathcal{L}_{\text{KD}}(f_S)$  (student better fits teacher on more unlabeled data) and  $C_g / \sqrt{n_u}$  decays, the RHS of Eq. (11) can drop below  $R(f_T)$ .
- **Optimization tolerance.**  $\varepsilon_{\text{opt}}$  captures suboptimality of empirical minimization; regularization / longer optimization reduces it.

### D.5.7 Concluding Remark

The preceding analysis bridges abstract statistical learning bounds with concrete, parameter-dependent quantities. By expressing the Rademacher complexities in terms of model smoothness and parameter norms, we obtain an explicit understanding of how both labeled and unlabeled data contribute to generalization.

In particular, the unified bound (11) highlights a clear trade-off: increasing model capacity or relaxing regularization tends to enlarge the complexity terms, whereas incorporating more unlabeled data, as captured by the distillation term, contributes to reducing the overall risk. This formal perspective complements the empirical observations reported in the main experiments.

Overall, these detailed derivations establish a transparent link between uniform convergence theory and the generalization behavior of our semi-supervised distillation framework. They further clarify how model regularity (e.g., spectral or Euclidean norms), loss smoothness, and the scale of unlabeled data jointly determine whether the student can outperform its teacher in expectation.

## E Discussion

**Limitations.** The proposed method has achieved excellent performance in extensive medical image segmentation tasks. However, factors such as complex anatomical structures, diverse imaging modes and inconsistent image quality in medical images may still pose significant challenges for segmentation. In such scenarios, suboptimal performance of our teacher SAM may further negatively impact the optimization of our efficient SAM backbone during semi-supervised learning. Therefore, future work will focus on enhancing the robustness of SAM in complex segmentation tasks and exploring effective strategies to reduce error propagation, thereby further improving segmentation accuracy and stability.

**Impacts.** This work holds significant implications for semi-supervised medical image segmentation. The proposed method improves segmentation accuracy even with limited annotated data and achieves high-quality results in complex medical scenarios. At the same time, it substantially reduces reliance on costly manual annotations while improving the efficiency of clinical diagnosis and treatment planning. Furthermore, our model contains less than 10% of the parameters of the original SAM, which significantly lowers deployment and maintenance costs and makes it more accessible to small and medium-sized medical institutions. Its simple and intuitive design also makes it easy for physicians to understand and use. However, the potential mis-segmentation risks for atypical cases still necessitate manual verification. Before widespread clinical application, its implementation must be accompanied by comprehensive regulatory frameworks and usage protocols to address associated liability and ethical concerns.

## References

- [1] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.
- [2] Chen Chen, Wenjia Bai, and Daniel Rueckert. Multi-task learning for left atrial segmentation on ge-mri. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9*, pages 292–301, 2019.
- [3] Spyridon (Spyros) Bakas. Brats miccai brain tumor dataset, 2020.
- [4] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In *MICCAI Workshop*, volume 5, page 12, 2015.
- [5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [6] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

- [7] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020.
- [8] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianying Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *MICCAI*, pages 318–329, 2021.
- [9] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *MICCAI*, pages 34–43, 2022.
- [10] Jinhua Liu, Christian Desrosiers, and Yuanfeng Zhou. Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In *MICCAI*, pages 140–150, 2022.
- [11] Zhenxi Zhang, Ran Ran, Chunna Tian, Heng Zhou, Xin Li, Fan Yang, and Zhicheng Jiao. Self-aware and cross-sample prototypical learning for semi-supervised medical image segmentation. In *MICCAI*, pages 192–201, 2023.
- [12] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *CVPR*, pages 11514–11524, 2023.
- [13] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *CVPR*, pages 4070–4080, 2024.
- [14] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.
- [15] Kevin Li and Pranav Rajpurkar. Adapting segment anything models to medical imaging via fine-tuning without domain pretraining. In *AAAI Workshop*, 2024.
- [16] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016.
- [17] Mary Phuong and Christoph H Lampert. Towards understanding knowledge distillation. In *ICML*, 2019.
- [18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [19] Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.
- [20] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin D Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 2018.
- [21] Colin Wei, Jason D. Lee, and Tengyu Ma. Theoretical analysis of semi-supervised learning via self-training. *Journal of Machine Learning Research*, 2021.
- [22] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation: A teacher-free framework. In *CVPR*, 2020.
- [23] Zhiqiang Xu and Ambuj Tewari. Understanding knowledge distillation in non-convex optimization. In *NeurIPS*, 2021.
- [24] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *NeurIPS*, 30, 2017.
- [25] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.
- [26] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *COLT*, 2018.