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

Haifeng Zhao[1,2], Haiyang Li[1,2], Lei-Lei Ma[1,2]*, Dengdi Sun[2,3]*
[1]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology, Anhui University, China
[2]Key Lab Intelligent Comp & Signal Proc, Minist Educ, Anhui University, China
[3]School of Artificial Intelligence, Anhui University, China

Abstract

The scarcity of annotated medical imaging data has driven significant progress in semi-supervised learning to alleviate reliance on expensive expert labeling. While foundational vision models such as the Segment Anything Model (SAM) exhibit robust generalization in generic segmentation tasks, their direct application to medical images often results in suboptimal performance. To address this challenge, in this work, we propose a novel fully SAM-based semi-supervised medical image segmentation framework and develop the corresponding knowledge distillation-based learning strategy. Specifically, we first employ an efficient SAM variant as the backbone network of the semi-supervised framework and update the default prompt embedding of SAM to unleash its full potential. Then, we utilize an original SAM, which is rich in prior knowledge, as the teacher to optimize our efficient student SAM backbone through hierarchical knowledge distillation and a dynamic loss weighting strategy. Extensive experiments on various medical datasets demonstrate that our method outperforms state-of-the-art semi-supervised segmentation approaches. Especially, our model requires less than 10% of the parameter size of the original SAM, enabling substantially lower deployment and storage overhead in real-world clinical settings.

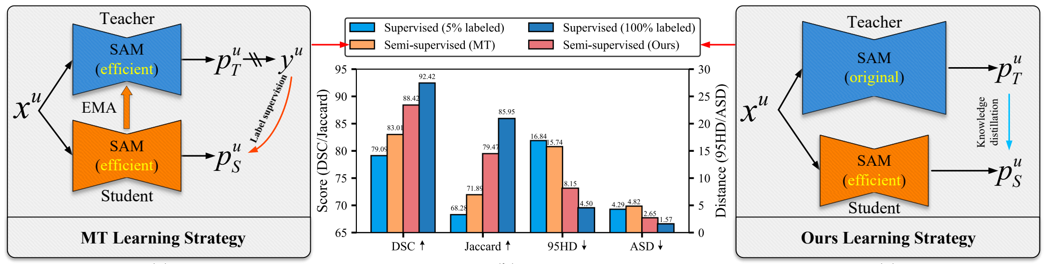
1 Introduction

Medical image segmentation aims to delineate precise anatomical structures from imaging data and serves as a fundamental basis for clinical applications [1, 2]. Recently, as an important foundational vision model for general image segmentation, the Segment Anything Model (SAM) [3] has been applied to medical images, resulting in various medical SAM variants. Through fine-tuning on medical datasets [4, 5, 6] and generating high-quality prompts [7, 8, 9, 10], these variants have achieved considerable segmentation results. Despite significant progress in medical image segmentation, the cumbersome and costly manual annotation process still hinders its development. To address this challenge, semi-supervised learning has gained traction as a robust approach that enhances model performance and generalization by leveraging both limited labeled data and abundant unlabeled data. However, there are still some key issues regarding how to effectively integrate the powerful capabilities of foundational models into semi-supervised medical image segmentation (SSMIS).

First, existing SSMIS frameworks have yet to fully leverage the capabilities of SAM. Lately, researchers have conducted numerous explorations into the application of SAM in SSMIS. For instance, SemiSAM [11] leverages coarse masks from the segmentation model to derive prompt points, which are then fed into SAM to generate more accurate pseudo-labels for model optimization. SFR [12] introduces an improved approach by pre-training SAM on labeled data, allowing it to

*Correspondence to: Lei-Lei Ma (xiaoleilei1990@gmail.com) and Dengdi Sun (sundengdi@163.com).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) (b) (c)

Figure 1: (a) Mean Teacher for ESDE. (b) Two semi-supervised approaches based on ESDE (Mean Teacher and our approach, colored in orange and red) with 5% labeled data and 95% unlabeled data vs. fully supervised learning (colored in blue) using 5% and 100% labeled data on the LA dataset. Our approach achieves competitive performance, with an 88.42% DSC, improving by 9.33% over the 5% labeled supervised setting. (c) Our proposed approach for ESDE.

directly generate high-quality pseudo-labels. However, due to the large parameter size of SAM, these approaches do not fully adopt SAM for semi-supervised segmentation but instead typically deploy it as an auxiliary component in SSMIS only for generating pseudo-labels as an additional source of supervision. This strategy inevitably restricts the potential of SAM. Moreover, such implementations are critically dependent on the prompts generated by the segmentation model, and inaccurate prompts can severely degrade the segmentation performance of SAM [13]. Therefore, the key challenge lies in maximizing the segmentation advantages of SAM while minimizing the impact of prompts, which is critical to advancing its application in SSMIS.

It is worth noting that to overcome the substantial computational demands of SAM, the community has proposed various lightweight approaches to improve efficiency without compromising accuracy. These methods focus on training lightweight models entirely from scratch [14, 15, 16] and use knowledge distillation with appropriate supervision for model training [17, 18, 19, 20, 21]. The success of efficient SAM variants inspires us to consider whether they can serve as backbone networks in semi-supervised learning frameworks, replacing traditional ConvNet-based backbones to fully leverage their capabilities. Moreover, when employing SAM as the segmentation model, updating its default prompt embedding can naturally enable high-quality segmentation results, while also mitigating the negative effects of inaccurate prompts. Therefore, we propose a novel fully SAM-based SSMIS framework that utilizes an **E**fficient **S**AM variant as the backbone with a **D**efault **E**mbedding (**ESDE**), which is promising to address the aforementioned challenges.

Second, commonly used learning strategies are incompatible with the fully SAM-based SSMIS framework. As shown in Figure 1 (a), we present a preliminary analysis by applying ESDE to a popular semi-supervised learning approach, Mean Teacher (MT) [22], where a teacher model is updated via exponential moving average (EMA) of a student model. However, as shown in Figure 1 (b), despite the availability of extensive unlabeled data, ESDE with MT achieves only a marginal improvement of less than 4% in Dice Similarity Coefficient (DSC). We attribute this limitation to the ViT-based architecture of SAM, which has weaker inductive biases compared to ConvNets, making it less effective in general semi-supervised paradigms [23]. Therefore, another key objective of this work is to develop a simple yet effective semi-supervised learning strategy for ESDE, in which the efficient SAM variant can better benefit from unlabeled data.

Based on the above analysis, as illustrated in Figure 1 (c), we propose a knowledge distillation-based learning strategy coupled with a customized training pipeline for ESDE to advance SSMIS. Specifically, we utilize an original SAM to optimize the efficient SAM backbone through knowledge distillation. The customized training process is divided into two steps: first, we apply parameter-efficient fine-tuning to the original SAM, progressively adapting its general segmentation capability to medical-specific tasks. Notably, this approach maintains the robustness of the foundational model while improving the precise recognition of anatomical structures. Then, the fine-tuned SAM serves as the teacher model, where it provides comprehensive guidance to the efficient student SAM through a hierarchical knowledge transfer mechanism operating on global contextual features and local boundary details. To address potential limitations, a dynamic weighting strategy adjusts the distillation intensity during training, thereby mitigating the excessive influence of the frozen teacher. Through the phased training pipeline, the framework effectively utilizes unlabeled data while preserving the segmentation strength of SAM in medical scenarios.

Overall, in this work, we design a fully SAM-based SSMIS framework and develop the corresponding knowledge distillation-based learning strategy. Our main contributions are summarized as follows:

- To our knowledge, we are the first to fully use efficient SAM as the backbone network in a semi-supervised learning framework. Meanwhile, we update the default prompt embedding of SAM to avoid the negative impact of inaccurate prompts. This innovative attempt harnesses its full power while reducing the complexity of deployment.
- We propose a novel semi-supervised learning strategy to address the limitations of ESDE in traditional settings. It constructs a hierarchical knowledge distillation pipeline to transfer structural knowledge and boundary details, and designs a dynamic weighting mechanism that adaptively adjusts distillation intensity based on the student's evolving capabilities.
- Comprehensive experiments across various medical segmentation tasks demonstrate that our proposed method achieves superior performance compared to other state-of-the-art semi-supervised methods.

2 Related Work

2.1 Semi-supervised Medical Image Segmentation

Since medical image annotation requires specialized expertise, incurs high costs, and is time-consuming, semi-supervised learning (SSL) [24, 25, 26] has emerged as an effective approach to tackle the challenge of insufficient labeled data in medical image segmentation. Most existing semi-supervised segmentation methods are typically grouped into two categories: consistency regularization [22] and pseudo-labeling [27]. Consistency regularization aims to learn more robust and generalizable representations by maintaining consistent predictions for the same input under different perturbations [28, 29]. Pseudo-labeling leverages the high-confidence predictions of the model on unlabeled data as temporary labels for supervised training [30, 31]. Unlike these methods, which are typically optimized using ConvNet-based backbones, we propose introducing an efficient SAM variant as the backbone network and designing a novel fully SAM-based framework.

2.2 SAM Adaptation for Medical Images

The Segment Anything Model (SAM) [3], trained on the large-scale SA-1B dataset, exhibits remarkable zero-shot generalization for natural images [32]. However, its performance drops sharply when applied to medical images [33, 34, 35], especially for unseen anatomical structures or pathologies [36, 37]. To better adapt SAM for medical images, researchers are focusing on generalizing SAM using automatic prompting techniques and different fine-tuning strategies. For instance, MaskSAM [38] provides accurate guidance to SAM through the design of a learnable prompt generator. Med-SA [5] fine-tunes the prompted SAM with points and boxes, incorporating lightweight adaptation blocks to extract domain-specific medical prior knowledge. Moreover, prompt-free SAM adaptation approaches introduced for medical segmentation suggest that prompts may not be completely essential [39, 40]. SAMed [6] integrates LoRA [41] layers into the image encoder while eliminating the prompts. Building on this progress, H-SAM [42] employs a hierarchical decoding structure to optimize the process of fine-tuning with limited medical data and attains impressive results. In this work, we also adopt a prompt-free strategy to decrease the adverse effects of inaccurate prompts, further optimizing the application of SAM in semi-supervised medical image segmentation.

2.3 Efficient SAM with Knowledge Distillation

Knowledge distillation (KD) is a model compression technique that improves a lightweight student by transferring knowledge from a complex teacher [43]. When deploying KD to accelerate SAM [3], the objective is to convey knowledge from the original, larger SAM to more compact and efficient SAM-like models. Considering the encoder-decoder architecture of SAM, KD strategies are typically divided into two approaches: distilling the entire SAM model or distilling only the image encoder. For example, MobileSAM [20] distills the image embedding from the image encoder into a lightweight ViT encoder while replicating the prompt-guided decoder. Based on this method, EfficientSAM [15] achieves a great speed-performance trade-off through masked image pretraining. However, TinySAM [17] points out that the absence of mask-level supervision for the student network

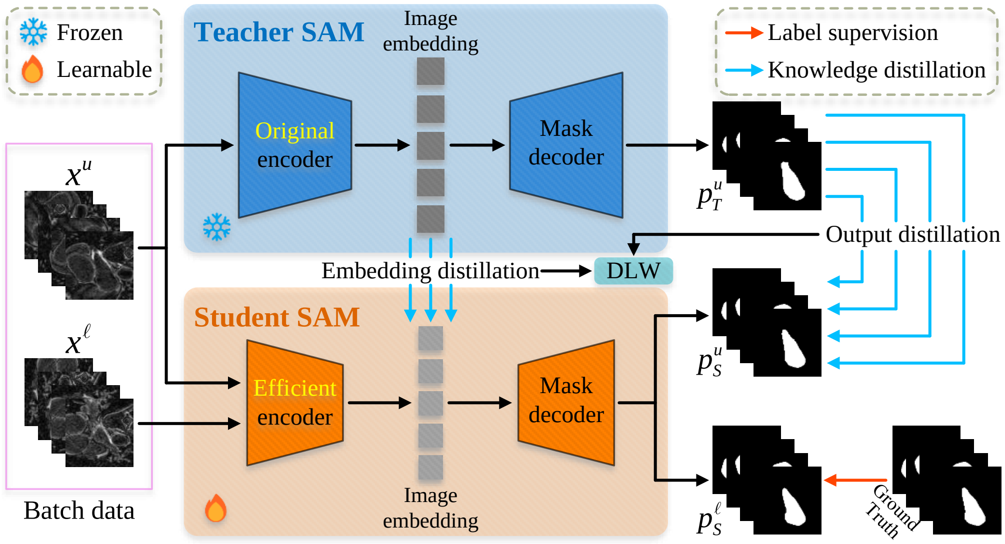


Figure 2: Overview of the proposed knowledge distillation-based semi-supervised learning strategy for ESDE. For labeled data, the student SAM receives direct supervision from ground-truth masks via segmentation losses. For unlabeled data, hierarchical distillation transfers knowledge from the frozen teacher (encoder embeddings and decoder outputs) to the student via KL divergence. Moreover, DLW strategy automatically adjusts the distillation intensity for better model optimization.

may lead to significant performance degradation and, therefore, proposes a full-stage distillation framework. Inspired by these methods, in this work, we leverage KD by transferring knowledge from the teacher model to refine the predictions on unlabeled samples during semi-supervised learning.

3 Preliminaries

Notations. Let $\mathcal{X} \subset \mathbb{R}^{H\times W}$ denote the image space, where each medical image $\boldsymbol{x} \in \mathcal{X}$ has a resolution of $H \times W$. The corresponding label space is defined as $\mathcal{Y} \subset \{1, \dots, C\}^{H\times W}$, where each pixel is assigned one of C anatomical or pathological categories. In the semi-supervised setting, we have two datasets:$\mathcal{D}^\ell = \{(\boldsymbol{x}_i^\ell, \boldsymbol{y}_i^\ell)\}_{i=1}^N, \mathcal{D}^u = \{\boldsymbol{x}_i^u\}_{i=N+1}^{N+M}$, where $\mathcal{D}^\ell$ contains N labeled samples and $\mathcal{D}^u$ contains M unlabeled samples, typically $N \ll M$. A SAM-based segmentation model $f_\Theta : \mathcal{X} \to \mathcal{Y}$ can be decomposed as an encoder-decoder:

$$f_\Theta(\boldsymbol{x}) = \mathcal{G}_{\Theta_g}\left(\mathcal{E}_{\Theta_e}(\boldsymbol{x}), \boldsymbol{q}\right) , \tag{1}$$

where $\mathcal{E}_{\Theta_e}$ denotes the image encoder, $\mathcal{G}_{\Theta_g}$ denotes the mask decoder, and $\boldsymbol{q}$ is the query of the mask decoder, which is formed by concatenating the prompt embedding with the output tokens. Let $\boldsymbol{\Theta}_f = \{\Theta^{(i)}\}_{i=1}^K$ be the set of all parameter tensors of model f_Θ. The total parameter count is defined as $P(f_\Theta) = \sum_{i=1}^K |\Theta^{(i)}|$, where $|\Theta^{(i)}|$ denotes the number of scalar parameters in tensor $\Theta^{(i)}$. In particular, if f_{Θ_1} and f_{Θ_2} are two models, $P(f_{\Theta_1}) > P(f_{\Theta_2})$ means f_{Θ_1} has more parameters.

Problem Definition. The goal of SSMIS is to learn a segmentation function f_Θ that approaches the performance of a fully supervised model using limited annotations. This can be formulated as a semi-supervised expected risk minimization problem:

$$\min_{\Theta} \mathbb{E}_{(\boldsymbol{x}^\ell,\boldsymbol{y}^\ell)\sim\mathcal{D}^\ell}[\mathcal{L}_{\text{sup}}(f_\Theta(\boldsymbol{x}^\ell), \boldsymbol{y}^\ell)] + \lambda\, \mathbb{E}_{\boldsymbol{x}^u\sim\mathcal{D}^u}[\mathcal{L}_{\text{unsup}}(f_\Theta, \boldsymbol{x}^u)] , \tag{2}$$

where $\mathcal{L}_{\text{sup}}$ is the supervised loss (e.g., Dice or cross-entropy), $\mathcal{L}_{\text{unsup}}$ exploits unlabeled data via consistency learning, pseudo-labeling, or knowledge distillation, and λ balances their contributions.

We adopt a teacher-student paradigm to tackle the above problem. The teacher model f_{Θ_T} provides supervision, while the student f_{Θ_S}, which has fewer parameters ($P(f_{\Theta_T}) > P(f_{\Theta_S})$), learns from labeled data and the teacher's guidance on unlabeled samples. Hierarchical knowledge distillation and dynamic loss weighting allow the student to inherit both the teacher's visual priors and domain-specific knowledge, yielding a high-performance yet lightweight segmentation model.

4 Method

4.1 Overview

As shown in Figure 2, our SSMIS framework employs a teacher model f_{Θ_T} (original SAM) and a student model f_{Θ_S} (efficient SAM variant), both using the default prompt embedding for segmentation. The training process comprises two sequential stages: *supervised fine-tuning* and *semi-supervised learning*. **In the first stage** (*supervised fine-tuning*), both teacher and student are fine-tuned on the labeled dataset $\mathcal{D}^\ell$. For the teacher model f_{Θ_T}, we apply the low-rank-based (LoRA) fine-tuning strategy to the image encoder, preserving generic visual features while incorporating medical domain knowledge. The student model f_{Θ_S} is fine-tuned simultaneously to learn task-specific representations with limited annotations. **In the second stage** (*semi-supervised learning*), the fine-tuned teacher is frozen and guides the student on unlabeled data $\mathcal{D}^u$. Crucially, hierarchical distillation pipeline transfers different-level feature representations from teacher to student, while dynamic loss weighting adaptively balances the distillation intensity, enabling efficient and stable knowledge transfer. **During testing**, only the lightweight student model f_{Θ_S} is used for inference, providing fast and memory-efficient predictions while maintaining high segmentation performance.

4.2 Fine-tuning Stage

As illustrated in Figure 3, our foundational teacher model is built upon the original SAM, and it is composed of three components: a ViT-based image encoder, a prompt encoder, and a mask decoder. The image encoder is the most heavyweight part, comprising a significant proportion of total parameters. Specifically, we freeze all layers in the image encoder and add a smaller, trainable bypass for each transformer block. These bypasses, which constitute the LoRA layers, first compress the transformer features into a low-rank space and subsequently reproject these condensed features to align with the channel dimensions of the output features in the frozen transformer blocks. During training, only the LoRA layers are updated, facilitating subtle yet impactful adjustments to the model and enabling efficient parameter adaptation with minimal memory overhead.

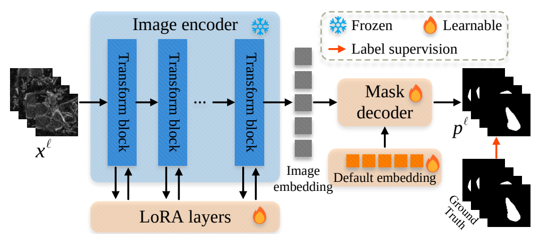


Figure 3: The pipeline for fine-tuning the teacher SAM. We freeze the image encoder and integrate LoRA layers into transformer blocks for efficient fine-tuning. Additionally, we update the prompt encoder using a default embedding to enable a prompt-free setting.

For the SSMIS task, the prompt encoder plays a crucial role in achieving more accurate segmentation results. Here, we update the default embedding, which refers to the embedding generated by the prompt encoder when no prompts are provided, to enhance SAM's performance and eliminate the need for manual inputs. This strategy not only avoids the potential pitfalls of inaccurate prompts, but also enables automated medical diagnosis. For the mask decoder, we update all parameters. In summary, we adopt the same implementation as SAMed [6] to fine-tune the teacher model. During fine-tuning, we utilize a supervised loss $\mathcal{L}_{\text{sup}}$ as the fine-tuning stage optimization objective $\mathcal{L}_{\text{ft}}$, enabling the model to learn accurate segmentation boundaries from labeled medical data effectively. The fine-tuning optimization is formulated as:

$$\mathcal{L}_{\text{ft}}(\boldsymbol{x}_i^\ell, \boldsymbol{y}_i^\ell) = \mathcal{L}_{\text{sup}}(\boldsymbol{p}_i^\ell, \boldsymbol{y}_i^\ell)\,, \quad \mathcal{L}_{\text{sup}}(\boldsymbol{p}_i^\ell, \boldsymbol{y}_i^\ell) = \lambda_{\text{dice}}\,\mathcal{L}_{\text{dice}}(\boldsymbol{p}_i^\ell, \boldsymbol{y}_i^\ell) + \lambda_{\text{ce}}\,\mathcal{L}_{\text{ce}}(\boldsymbol{p}_i^\ell, \boldsymbol{y}_i^\ell)\,, \tag{3}$$

where $\boldsymbol{p}_i^\ell = f_\Theta(\boldsymbol{x}_i^\ell)$ denotes the model prediction for a labeled image, and $\boldsymbol{y}_i^\ell$ is the corresponding ground-truth mask. $\mathcal{L}_{\text{dice}}$ and $\mathcal{L}_{\text{ce}}$ represent the Dice loss and Cross-entropy loss, respectively, with λ_{dice} and λ_{ce} controlling their relative contributions.

To obtain an efficient SAM backbone, we replace the image encoder of the original SAM with a much smaller ViT, while the rest of the components remain unchanged. For this student SAM, similarly, we also retain the strategy of updating the default embedding. During fine-tuning, all parameters are optimized, utilizing the same loss function as the teacher model.

4.3 Semi-supervised Learning Stage

After fine-tuning, we further optimize the model through semi-supervised learning. As shown in Figure 2, our framework employs a teacher-student architecture to strategically leverage both labeled and unlabeled data. Specifically, the teacher model remains frozen to preserve reliable knowledge and generalization capability, while the student model maintains full trainability to assimilate two learning signals: direct supervision from labeled data via ground-truth annotations and knowledge distillation from the teacher's embedding-level and logit-level for unlabeled data. In this work, the proposed distillation mechanism is implemented through the newly designed hierarchical distillation pipeline and dynamic loss weighting strategy, both of which are detailed subsequently.

Hierarchical Knowledge Distillation. Inspired by MobileSAM [20], we adopt the image encoder output as the distilled information. Since $\mathcal{E}_T(\boldsymbol{x}_i)$ and $\mathcal{E}_S(\boldsymbol{x}_i)$ are feature maps of shape $D \times H' \times W'$, we compute KL divergence for each spatial position after softmax along the channel dimension. Given a batch of N' unlabeled images $\{\boldsymbol{x}_i\}_{i=1}^{N'} \subset \mathcal{D}^u$, the embedding-level distillation loss is:

$$\mathcal{L}_{\text{emb}} = \frac{1}{N' \cdot H' \cdot W'} \sum_{i=1}^{N'} \sum_{u=1}^{H'} \sum_{v=1}^{W'} D_{\text{KL}}\left(\hat{\mathcal{E}}_T(\boldsymbol{x}_i)_{:,u,v} \parallel \hat{\mathcal{E}}_S(\boldsymbol{x}_i)_{:,u,v}\right), \tag{4}$$

where $\hat{\mathcal{E}}_T(\boldsymbol{x}_i)_{:,u,v} = \text{softmax}(\mathcal{E}_T(\boldsymbol{x}_i)_{:,u,v})$, $\hat{\mathcal{E}}_S(\boldsymbol{x}_i)_{:,u,v} = \text{softmax}(\mathcal{E}_S(\boldsymbol{x}_i)_{:,u,v})$, $D_{\text{KL}}(p\|q) = \sum_{d=1}^{D} p_d \log \frac{p_d}{q_d}$, with u, v indexing spatial positions and d denoting the number of dimensions.

Although the global features extracted by the image encoder provide a broad understanding of the image structure, they are insufficient for the precise requirements of pixel-level mask prediction. Segmentation tasks rely more on features near the output layer, which capture localized details and boundaries. Hence, the teacher's logit output is also selected as a distillation target to provide fine-grained guidance. The logit-level distillation loss is defined as:

$$\mathcal{L}_{\text{logit}} = \frac{1}{H \cdot W} \sum_{u=1}^{H} \sum_{v=1}^{W} D_{\text{KL}}\left(\hat{\mathcal{G}}_T(\boldsymbol{x}_i)_{:,u,v} \parallel \hat{\mathcal{G}}_S(\boldsymbol{x}_i)_{:,u,v}\right), \tag{5}$$

where $\hat{\mathcal{G}}_T(\boldsymbol{x}_i)_{:,u,v} = \text{softmax}\big(\mathcal{G}_T(\mathcal{E}_T(\boldsymbol{x}_i), \boldsymbol{q}_T)_{:,u,v}\big)$, $\hat{\mathcal{G}}_S(\boldsymbol{x}_i)_{:,u,v} = \text{softmax}\big(\mathcal{G}_S(\mathcal{E}_S(\boldsymbol{x}_i), \boldsymbol{q}_S)_{:,u,v}\big)$.

Combining the embedding and logit distillation losses helps the student better replicate the teacher's performance. Thus, the overall knowledge distillation loss function is given as:

$$\mathcal{L}_{\text{kd}} = \lambda_{\text{emb}} \mathcal{L}_{\text{emb}} + \lambda_{\text{logit}} \mathcal{L}_{\text{logit}}, \tag{6}$$

where λ_{emb} and λ_{logit} are weight coefficients that balance the contributions of the embedding and logit distillation losses, respectively.

Dynamic Loss Weighting. To prevent the static knowledge of the frozen teacher from dominating the learning process, we design a dynamic loss weighting (DLW) strategy that automatically adjusts the distillation intensity based on comparative performance metrics. The key mechanism involves monitoring the supervised losses of both models on labeled data: the teacher's loss $\mathcal{L}_{\text{sup}}^T$ serves as a fixed reference benchmark after fine-tuning, while the student's loss $\mathcal{L}_{\text{sup}}^S$ reflects its current learning progress. At each epoch t, we update the distillation weight λ_{kd} through conditional decay:

$$\lambda_{\text{kd}}^{(t)} = \begin{cases} \alpha \cdot \lambda_{\text{kd}}^{(t-1)}, & \mathcal{L}_{\text{sup}}^S < \mathcal{L}_{\text{sup}}^T, \\ \lambda_{\text{kd}}^{(t-1)}, & \text{otherwise}, \end{cases} \tag{7}$$

where α is a scaling factor (e.g., 0.95) for gradual decay and $\lambda_{\text{kd}}^{(0)} = 1.0$ initiates strong guidance. Note that $\mathcal{L}_{\text{sup}}^T$ is only used to compute λ_{kd} and does not contribute to the backpropagation of gradients. DLW helps the student rely more on the teacher's guidance in the early stages while gradually focusing on self-learning as training progresses, which aids in adapting to unlabeled data and improving generalization to unseen data.

In summary, the overall optimization objective for semi-supervised learning is written as:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{kd}} \mathcal{L}_{\text{kd}}. \tag{8}$$

Differing from $\mathcal{L}_{\text{sup}}^S$, which denotes the supervised loss of the student model on labeled data summed over an entire epoch, $\mathcal{L}_{\text{sup}}$ here denotes the loss for each training iteration.

Table 1: Comparisons with SOTA semi-supervised segmentation methods on the LA dataset.

Method	Scans used		Metrics				Scans used		Metrics			
	Labeled	Unlabeled	DSC↑	Jaccard↑	95HD↓	ASD↓	Labeled	Unlabeled	DSC↑	Jaccard↑	95HD↓	ASD↓
UA-MT [29]	4(5%)	76(95%)	82.26	70.98	13.71	3.82	8(10%)	72(90%)	87.79	78.39	8.68	2.12
SASSNet [49]			81.60	69.63	16.16	3.58			87.54	78.05	9.84	2.59
DTC [50]			81.25	69.33	14.90	3.99			87.51	78.17	8.23	2.36
URPC [51]			82.48	71.35	14.65	3.65			86.92	77.03	11.13	2.28
MC-Net [52]			83.59	72.36	14.07	2.70			87.62	78.25	10.03	1.82
SS-Net [31]			86.33	76.15	9.97	2.31			88.55	79.62	7.49	1.90
BCP [30]			88.02	78.72	**7.90**	**2.15**			89.62	81.31	6.81	**1.76**
Ours			**88.42**	**79.47**	8.15	2.65			**90.47**	**82.71**	**6.41**	1.97

Table 2: Comparisons with SOTA semi-supervised segmentation methods on the Brats-2019 dataset.

Method	Scans used		Metrics				Scans used		Metrics			
	Labeled	Unlabeled	DSC↑	Jaccard↑	95HD↓	ASD↓	Labeled	Unlabeled	DSC↑	Jaccard↑	95HD↓	ASD↓
DAN [53]	25(10%)	225(90%)	81.71	71.43	15.15	2.32	50(20%)	200(80%)	83.31	73.53	10.86	2.23
DTC [50]			81.75	71.63	15.73	2.56			83.43	73.56	14.77	2.34
CPS [54]			82.52	72.66	13.08	2.66			84.01	74.02	12.16	2.18
URPC [51]			82.59	72.11	13.88	3.72			82.93	72.57	15.93	4.19
CPCL [55]			83.36	73.23	11.74	1.99			83.48	74.08	9.53	2.08
AC-MT [56]			83.77	73.96	11.37	**1.93**			84.63	74.39	9.50	2.11
MLRPL [57]			84.29	74.74	9.57	2.55			85.47	76.32	7.76	**2.00**
Ours			**85.14**	**75.56**	**8.58**	2.57			**86.46**	**77.21**	**7.64**	2.21

5 Experiments

We validate our proposed method on three widely-used semi-supervised medical image segmentation datasets: the LA dataset [44], the Brats-2019 dataset [45], and the PROMISE12 dataset [46]. Additionally, to facilitate a comprehensive comparison with existing prompt-free medical SAM variants, we conduct experiments on the Synapse Multi-Organ CT dataset [47]. For the foundational teacher SAM, we conduct all experiments based on the "ViT-B" version, while for the efficient student SAM, we replace the original image encoder with TinyViT-5M [48]. More details of the datasets, evaluation metrics, and method implementation, as well as additional experimental results such as the effectiveness of the fine-tuned teacher and model complexity analysis, are provided in the Appendix.

5.1 Comparison with Sate-of-the-Art Methods

Results on LA Dataset. We evaluate our method on the LA dataset against state-of-the-art semi-supervised methods, including UA-MT [29], SASSNet [49], DTC [50], URPC [51], MC-Net [52], SS-Net [31], and BCP [30], using labeled ratios of 5% and 10%. As shown in Table 1, our approach outperforms these competitors, achieving a 0.40% improvement in DSC with 5% labeled data and 0.85% with 10%. Additionally, other metrics highlight its competitiveness, with gains of 1.40% in Jaccard and 0.40 in 95HD under the 10% setting. It demonstrates the feasibility of using SAM as the backbone network with a default embedding and validates the effectiveness of the proposed knowledge distillation-based semi-supervised learning strategy. Moreover, qualitative results in Figure 4 demonstrate that our method produces finer segmentation boundaries and better overall agreement with ground-truth annotations compared to existing approaches.

Results on Brats-2019 Dataset. Brain tumor segmentation is highly challenging due to variations in tumor appearance and uncertain boundaries. To demonstrate the effectiveness of our approach, we conduct experiments on the Brats-2019 dataset with 10% and 20% labeled ratio, comparing it with several methods, including DAN [53], DTC [50], CPS [54], URPC [51], CPCL [55], AC-MT [56], and MLRPL [57]. As evidenced in Table 2, our method achieves state-of-the-art performance on both labeled protocols. With 10% labeled data, we obtain 85.14% DSC and 75.56% Jaccard scores, outperforming the strongest baseline (MLRPL) by 0.85% and 0.82%, respectively. This advantage further expands to 1.0% DSC and 0.89% Jaccard improvements under the 20% labeled setting. Furthermore, our approach reduces 95HD by 0.99 compared to MLRPL with 10% labeled data, demonstrating a superior boundary adherence capability for complex tumor structures. The consistent

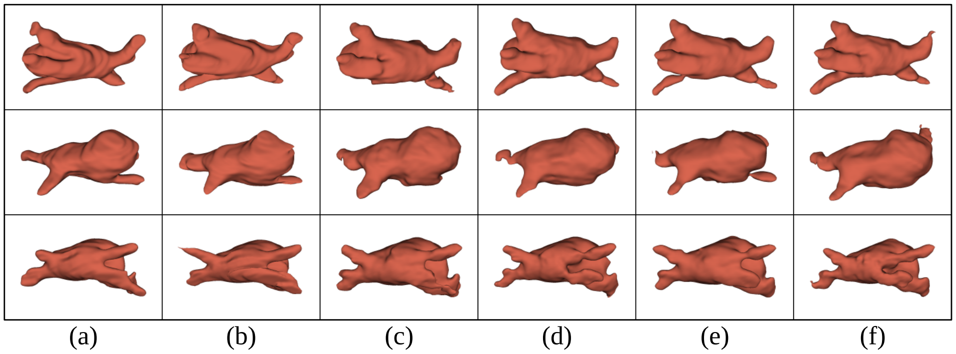


Figure 4: Visualization of segmentation results on the LA dataset with 10% labeled data. (a) Groud-truth. (b) Ours results. (c) BCP results. (d) SS-Net results. (e) MC-Net results. (f) DTC results.

performance gains validate the effectiveness of our proposed hierarchical knowledge distillation and dynamic loss weighting strategy in handling ambiguous tumor margins.

Results on PROMISE12 Dataset. We also perform experiments on the PROMISE12 dataset with 20% labeled ratio against CCT [58], URPC [51], SS-Net [31], SLC-Net [59], SCP-Net [60], BCP [30], and ABD [28], as well as ABD with 10% labeled ratio. Detailed results are shown in the Appendix.

5.2 Ablation Studies

We conduct ablation studies to evaluate the impact of key components in our method, including combinations of knowledge distillation losses, the ratios of λ_{emb} and λ_{logit} (Eq (6)), fine-tuning the student model, and the proposed DLW strategy. All experiments here are performed on the LA dataset, with ablation results on other datasets provided in the Appendix.

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

Embedding loss	Logit loss	Scans used: Labeled	Unlabeled	Metrics: DSC↑	Jaccard↑	95HD↓	ASD↓
✓		4(5%)	76(95%)	81.72	70.92	23.24	6.24
	✓			88.07	79.00	10.73	2.82
✓	✓			**88.42**	**79.47**	**8.15**	**2.65**
✓		8(10%)	72(90%)	84.53	75.18	11.75	3.57
	✓			90.19	82.26	6.64	2.02
✓	✓			**90.47**	**82.71**	**6.41**	**1.97**

Table 4: Ablation study on ratios of λ_{emb} and λ_{logit} with 10% labeled data.

λ_{emb}	λ_{logit}	Metrics: DSC↑	Jaccard↑	95HD↓	ASD↓
1/4	3/4	90.28	82.39	6.70	2.03
1/3	2/3	**90.47**	**82.71**	**6.41**	**1.97**
1/2	1/2	90.25	82.35	6.81	2.13
2/3	1/3	89.92	81.86	6.95	2.14
3/4	1/4	89.38	80.94	9.24	2.94

Table 5: Effectiveness of fine-tuning in the student model.

Method	Scans used: Labeled	Unlabeled	Metrics: DSC↑	Jaccard↑	95HD↓	ASD↓
w/o fine-tune	4(5%)	76(95%)	86.57	76.73	9.95	3.08
w/ fine-tune			**88.42**	**79.47**	**8.15**	**2.65**
w/o fine-tune	8(10%)	72(90%)	88.90	80.19	7.05	2.30
w/ fine-tune			**90.47**	**82.71**	**6.41**	**1.97**

Table 6: Effectiveness of the proposed DLW strategy.

Method	Scans used: Labeled	Unlabeled	Metrics: DSC↑	Jaccard↑	95HD↓	ASD↓
w/o DLW	4(5%)	76(95%)	88.00	78.94	**8.10**	2.67
w/ DLW			**88.42**	**79.47**	8.15	**2.65**
w/o DLW	8(10%)	72(90%)	90.02	81.97	7.04	2.25
w/ DLW			**90.47**	**82.71**	**6.41**	**1.97**

Different Knowledge Distillation Losses. Table 3 examines the effectiveness of different combinations of knowledge distillation losses, including embedding loss and logit loss, to evaluate their contributions. The results reveal that the logit loss plays a critical role, as it closely aligns with the supervised information and directly impacts the evaluation metrics. Using the embedding loss alone shows weaker performance, while the logit loss alone yields noticeable improvements. However, combining both losses significantly enhances segmentation performance. These findings highlight the complementary roles of embedding loss, which enables the student encoder to learn semantic knowledge of image-dense features and spatially structured relationships from the teacher encoder, and logit loss, which leverages the supervised alignment output to better align with the teacher.

Effect of Balancing Embedding and Prediction Losses. To explore the impact of different weight ratios between embedding loss and logit loss, we conduct an ablation study with 10% labeled data, as

Table 7: Comparisons with SAM-assisted semi-supervised segmentation methods on the LA dataset.

Method	Scans used		Metrics				Scans used		Metrics			
	Labeled	Unlabeled	DSC↑	Jaccard↑	95HD↓	ASD↓	Labeled	Unlabeled	DSC↑	Jaccard↑	95HD↓	ASD↓
SemiSAM [11]	4(5%)	76(95%)	80.42	68.05	18.23	5.16	8(10%)	72(90%)	84.45	73.75	14.56	3.31
UP-SAM [62]			84.06	72.90	13.78	3.14			-	-	-	-
SFR [12]			87.95	78.83	9.23	2.89			89.99	81.93	7.07	2.04
Ours			**88.42**	**79.47**	**8.15**	**2.65**			**90.47**	**82.71**	**6.41**	**1.97**

Table 8: Comparisons with SOTA prompt-free medical SAM variants on the Synapse dataset with 10% labeled data.

Method	Params	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach	Metrics	
										DSC↑	95HD↓
SAM Adapter [5]	131.5M	66.74	22.38	66.77	68.38	89.69	26.76	72.42	53.15	58.29	54.22
AutoSAM [40]	112.5M	75.19	24.87	76.53	77.44	88.06	34.58	68.80	52.70	62.27	31.67
SAMed [6]	108.8M	78.72	63.15	82.62	**82.25**	92.72	52.12	85.82	67.20	75.58	23.02
H-SAM [42]	112.3M	79.65	59.76	82.71	82.14	91.73	47.48	**86.35**	**75.31**	75.68	21.34
Ours	10.1M	**86.58**	**67.22**	**83.23**	79.04	**93.00**	**57.08**	86.18	75.05	**78.42**	**17.87**

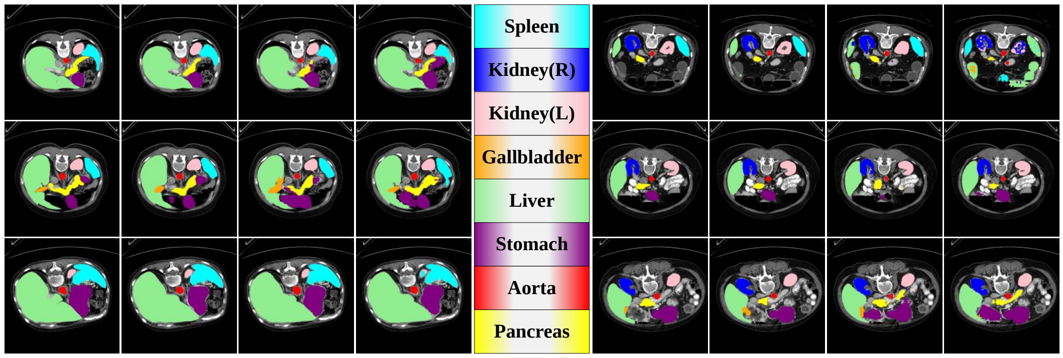
shown in Table 4. The results indicate that the balance between the two losses significantly affects segmentation performance. When the ratio $\lambda_{\mathrm{emb}} : \lambda_{\mathrm{logit}}$ is set to $1/3 : 2/3$, the best performance is achieved. However, as λ_{emb} increases, a noticeable decline in performance is observed, with DSC dropping from 90.47% to 89.38% at $3/4 : 1/4$. This indicates that overemphasizing the embedding loss undermines the contribution of the logit loss, which is more aligned with the supervised signals.

Fine-tuning in Student. Motivated by [61], we apply fine-tuning to the student model prior to semi-supervised learning. This step helps the student better adapt to the target task by leveraging labeled data for initial optimization, creating a more robust foundation for knowledge distillation. As shown in Table 5, fine-tuning reliably improves performance across all metrics, with significant gains observed in DSC and Jaccard scores. These results highlight the importance of initializing the student model with task-specific knowledge, which enhances its capacity to benefit from the subsequent semi-supervised learning process.

Dynamic Loss Weighting Strategy. In this work, we propose the DLW strategy to mitigate the potential over-reliance of the student model on the fixed outputs of the teacher model. To validate its effectiveness, we perform ablation experiments to evaluate its impact on segmentation performance. As shown in Table 6, incorporating DLW consistently enhances the results. For 5% labeled data, DLW improves the DSC to 88.42% and slightly reduces the ASD to 2.65. Similarly, for 10% labeled data, it significantly decreases the ASD from 2.25 to 1.97. By dynamically adjusting the weight of the distillation loss during training, DLW allows the student model to initially leverage the teacher's guidance while gradually shifting focus toward self-learning. This adaptive mechanism supports effective utilization of unlabeled data and contributes to better generalization.

5.3 Comparisons with SAM-assisted Semi-supervised Methods

In this work, we directly use SAM as the backbone network with a default embedding and design a novel knowledge distillation-based learning strategy. Therefore, we also conduct quantitative experiments on the LA dataset to compare the proposed method with three existing SAM-assisted semi-supervised methods: SemiSAM [11], UP-SAM [62], and SFR [12]. It is important to highlight that these methods use SAM as an auxiliary component to a ConvNet backbone for semi-supervised learning, and the outputs from the ConvNet are regarded as the final segmentation maps. As shown in Table 7, our method achieves the best performance, verifying that employing SAM as the segmentation pipeline to unleash its capability yields more substantial benefits than using it as an auxiliary module. This superiority can be attributed to the fact that the ConvNet-centric architectures of existing methods and their provision of unreliable prompts inevitably compromise the inherent segmentation potential of SAM, while our model with prompt-free setting successfully circumvents error accumulation in pseudo-label generation.



(a) (b) (c) (d) (a) (b) (c) (d)

Figure 5: Visualization of segmentation results on the Synapse dataset with 10% labeled data. (a) Groud-truth. (b) Ours results. (c) H-SAM results. (d) SAMed results.

5.4 Comparisons with Prompt-free Medical SAM Variants

For a more comprehensive evaluation, we further compare our method with several prompt-free medical SAM variants, including SAM Adapter [5], AutoSAM [40], SAMed [6], and H-SAM [42], all of which default to the "ViT-B" version. Table 8 shows the quantitative results on the multi-class task using the Synapse dataset with 10% labeled data. For fair comparison, we upsample 224×224 CT images to 512×512 as input instead of directly using a resolution of 512×512, to maintain consistency with the methods mentioned above. Notably, compared to these fully supervised methods, our method compresses the knowledge of a teacher SAM (equivalent to SAMed) into a student SAM with more than $10 \times$ fewer parameters, while also effectively leveraging a large amount of unlabeled data. It can be observed that our method consistently outperforms the other methods across the majority of organs, achieving higher segmentation accuracy while significantly reducing inference costs. Moreover, Figure 5 gives some qualitative results where our model yields smoother and more accurate segmentation regions compared to other methods.

6 Conclusion

In this work, we explore a better adaptation of SAM in semi-supervised medical image segmentation. Capitalizing on advancements in lightweight SAM techniques, we pioneer its deployment as the backbone network in semi-supervised frameworks to fully leverage its segmentation capability. To address the incompatibility of conventional semi-supervised methods with the SAM backbone, we develop a novel knowledge distillation-based learning strategy that achieves hierarchical distillation by aligning the anatomical semantics of the encoder with the boundary details of the decoder, and incorporates dynamic loss weighting to progressively reduce the distillation intensity for better exploitation of unlabeled data. In this way, the proposed method achieves higher segmentation accuracy while significantly reducing model parameters, enhancing the feasibility of clinical deployment. This work also establishes a new technical paradigm for the practical implementation of foundational models in medical image segmentation.

Acknowledgements

This project is funded by the National Natural Science Foundation of China (62576006, 62076005, 62472004), Provincial Quality Project of Education in the New Era in 2023 (Postgraduate Education 2023lhpysfjd009), and the University Synergy Innovation Program of Anhui Province, China (GXXT-2021-002, GXXT-2022-029). We sincerely thank Professor Bin Luo for his generous support of this project. We are grateful to Ke Zhu (LAMDA, Nanjing University) and Ming-Kun Xie (RIKEN AIP) for their valuable suggestions. We also thank the reviewers for their comments and suggestions. Finally, we acknowledge the High-Performance Computing Platform of Anhui University for providing computational resources for this project.

References

[1] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[2] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023.

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

[5] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.

[6] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[7] Zhen Chen, Qing Xu, Xinyu Liu, and Yixuan Yuan. Un-sam: Universal prompt-free segmentation for generalized nuclei images. *arXiv preprint arXiv:2402.16663*, 2024.

[8] Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In *MICCAI*, pages 24–34, 2024.

[9] Qi Wu, Yuyao Zhang, and Marawan Elbatel. Self-prompting large vision models for few-shot medical image segmentation. In *MICCAI Workshop*, pages 156–167, 2023.

[10] Chunpeng Zhou, Kangjie Ning, Qianqian Shen, Sheng Zhou, Zhi Yu, and Haishuai Wang. Sam-sp: Self-prompting makes sam great again. *arXiv preprint arXiv:2408.12364*, 2024.

[11] Yichi Zhang, Yuan Cheng, and Yuan Qi. Semisam: Exploring sam for enhancing semi-supervised medical image segmentation with extremely limited annotations. *arXiv preprint arXiv:2312.06316*, 2023.

[12] Shumeng Li, Lei Qi, Qian Yu, Jing Huo, Yinghuan Shi, and Yang Gao. Stitching, fine-tuning, re-training: A sam-enabled framework for semi-supervised 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

[13] 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.

[14] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[15] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *CVPR*, pages 16111–16121, 2024.

[16] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

[17] Han Shu, Wenshuo Li, Yehui Tang, Yiman Zhang, Yihao Chen, Houqiang Li, Yunhe Wang, and Xinghao Chen. Tinysam: Pushing the envelope for efficient segment anything model. In *AAAI*, volume 39, pages 20470–20478, 2025.

[18] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *CVPR*, pages 15909–15920, 2024.

[19] Haobo Yuan, Xiangtai Li, Lu Qi, Tao Zhang, Ming-Hsuan Yang, Shuicheng Yan, and Chen Change Loy. Mamba or rwkv: Exploring high-quality and high-efficiency segment anything model. *arXiv preprint arXiv:2406.19369*, 2024.

[20] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[21] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2312.06660*, 2023.

[22] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30, 2017.

[23] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Semi-supervised vision transformers. In *ECCV*, pages 605–620, 2022.

[24] Ming-Kun Xie, Jiahao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. *NeurIPS*, 36:25731–25747, 2023.

[25] Jia-Hao Xiao, Ming-Kun Xie, Heng-Bo Fan, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Dual-decoupling learning and metric-adaptive thresholding for semi-supervised multi-label learning. In *ECCV*, pages 437–454, 2024.

[26] Hao-Zhe Liu, Ming-Kun Xie, Chen-Chen Zong, and Sheng-Jun Huang. Asymmetric beta loss for evidence-based safe semi-supervised multi-label learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1909–1920, 2024.

[27] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, volume 3, page 896, 2013.

[28] Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *CVPR*, pages 4070–4080, 2024.

[29] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, pages 605–613, 2019.

[30] 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.

[31] 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.

[32] Mingjin Zhang, Qian Xu, Yuchun Wang, Xi Li, and Haojuan Yuan. Mirsam: multimodal vision-language segment anything model for infrared small target detection. *Visual Intelligence*, 3(1):1–13, 2025.

[33] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023.

[34] Chuanfei Hu, Tianyi Xia, Shenghong Ju, and Xinde Li. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. *arXiv preprint arXiv:2304.08506*, 2023.

[35] Tassilo Wald, Saikat Roy, Gregor Koehler, Nico Disch, Maximilian Rouven Rokuss, Julius Holzschuh, David Zimmerer, and Klaus Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. In *Medical Imaging with Deep Learning, short paper track*, 2023.

[36] Sovesh Mohapatra, Advait Gosai, and Gottfried Schlaug. Sam vs bet: A comparative study for brain extraction and segmentation of magnetic resonance images using deep learning. *arXiv preprint arXiv:2304.04738*, 2023.

[37] An Wang, Mobarakol Islam, Mengya Xu, Yang Zhang, and Hongliang Ren. Sam meets robotic surgery: an empirical study on generalization, robustness and adaptation. In *MICCAI*, pages 234–244, 2023.

[38] Bin Xie, Hao Tang, Bin Duan, Dawen Cai, and Yan Yan. Masksam: Towards auto-prompt sam with mask classification for medical image segmentation. *arXiv preprint arXiv:2403.14103*, 2024.

[39] Weijia Feng, Lingting Zhu, and Lequan Yu. Cheap lunch for medical image segmentation by fine-tuning sam on few exemplars. In *MICCAI Workshop*, pages 13–22, 2023.

[40] Xinrong Hu, Xiaowei Xu, and Yiyu Shi. How to efficiently adapt large segmentation model (sam) to medical images. *arXiv preprint arXiv:2306.13731*, 2023.

[41] 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.

[42] Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, and Yuyin Zhou. Unleashing the potential of sam for medical adaptation via hierarchical decoding. In *CVPR*, pages 3511–3522, 2024.

[43] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS Workshop*, 2015.

[44] 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.

[45] Spyridon (Spyros) Bakas. Brats miccai brain tumor dataset, 2020.

[46] 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.

[47] 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.

[48] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *ECCV*, pages 68–85, 2022.

[49] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *MICCAI*, pages 552–561, 2020.

[50] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *AAAI*, volume 35, pages 8801–8809, 2021.

[51] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong 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.

[52] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *MICCAI*, pages 297–306, 2021.

[53] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *MICCAI*, pages 408–416, 2017.

[54] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021.

[55] Zhe Xu, Yixin Wang, Donghuan Lu, Lequan Yu, Jiangpeng Yan, Jie Luo, Kai Ma, Yefeng Zheng, and Raymond Kai-yu Tong. All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3174–3184, 2022.

[56] Zhe Xu, Yixin Wang, Donghuan Lu, Xiangde Luo, Jiangpeng Yan, Yefeng Zheng, and Raymond Kai-yu Tong. Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. *Medical Image Analysis*, 88:102880, 2023.

[57] Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, and Shaozi Li. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Medical Image Analysis*, 94:103111, 2024.

[58] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020.

[59] 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.

[60] 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.

[61] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. *NeurIPS*, 35:25697–25710, 2022.

[62] Wenjing Lu, Yi Hong, and Yang Yang. Up-sam: Uncertainty-informed adaptation of segment anything model for semi-supervised medical image segmentation. In *BIBM*, pages 2256–2261, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope, clearly outlining both the methodological innovations and empirical findings. Subsequent sections provide firm theoretical grounding and robust experimental validation for each claim.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The information needed to reproduce the main experimental results is introduced in Section 4, Section 5, and the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We are currently not providing open access to the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the training and test details can be seen in Section 5 and the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our experiment did not include error bars, confidence intervals, or other statistical significance tests.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sufficient information on the computer resources can be seen in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper fully compliant with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both the potential positive societal impacts and negative societal impacts of the work in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.