A Ablation study on SAM fine-tuning

To evaluate the impact of fine-tuning the SAM encoder in our framework using box-level supervision, we conducted an ablation study. The results, presented in Table A, indicate a reduction in segmentation performance when the SAM encoder is fine-tuned. This reduction can be attributed to the coarse-grained nature of the box-level annotations, which contrasts with the typical fine-tuning practice where annotations are more precise than those used in pre-training. Additionally, fine-tuning the SAM encoder required significantly more GPU memory and increased training time.

Table A: Performance comparison of SAM Encoder: ***Fine-tuning vs. Frozen***.

	SAM Encoder	mIoU (%)	mACC (%)	training time
WPS-SAM	fine-tuning	65.16	75.79	127h
	frozen	**68.93**	**79.53**	59h

B Ablation study on network architectures

To investigate the impact of different network architectures, we conducted an ablation study comparing classical CNN architectures within an anchor-based framework to a query-based transformer architecture. The results, presented in Table B, indicate a performance decline when using the anchor-based CNN approach. This decline may be attributed to the incompatibility between the CNNs and the transformer architectures when combined within the model.

Table B: Performance comparison of different network architectures.

	architecture	mIoU (%)	mACC (%)
Student prompter	anchor-based (CNN)	57.23	68.11
	query-based (Transformer)	**68.93**	**79.53**

C Ablation study on hyper-parameters

We conducted detailed ablation studies on critical hyper-parameters to understand their impact on performance. The results, presented in Table C, provide insights into how different hyper-parameter settings affect our framework.

Table C: Performance comparison of different hyper-parameters.

	α	β	λ_{cls}	λ_{reg}	mIoU (%)	mACC (%)
WPS-SAM	1	1	1	1	39.35	47.40
	5	1	5	1	56.94	69.59
	5	1	10	1	61.93	77.07
	5	10	10	1	67.24	78.49
	5	**20**	**10**	**1**	**68.93**	**79.53**
	5	50	10	1	67.82	77.54