Learning from Noisy Prompts: Saliency-Guided Prompt Distillation for Robust Segmentation with SAM

Supplementary Material

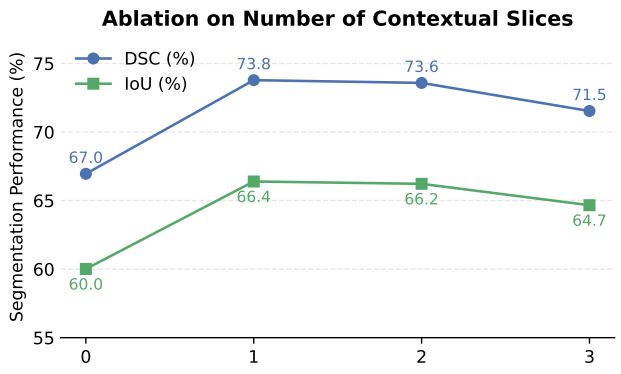

Figure 1. **Ablation on the number of contextual slices** n in CPD on the TI dataset.

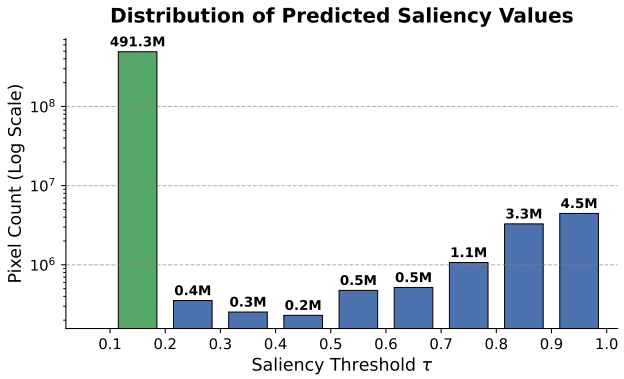

Figure 2. **Distribution of predicted saliency values on the TI dataset.** The histogram is plotted on a logarithmic scale to visualize the full range of pixel counts. It reveals a highly skewed distribution dominated by a massive peak of high-confidence background predictions (491.3M pixels). In contrast, all potential foreground signals (values > 0.2) collectively account for less than 2% of the total pixels, highlighting their extreme sparsity. This clear separation between confident background and sparse foreground explains the model's robustness to the saliency threshold τ.

1. Implementation Details

We implement SPD using the PyTorch framework and conduct all experiments on 8 NVIDIA RTX 4090 GPUs, with a per-GPU batch size of 4. The model is trained using the Adam optimizer with an initial learning rate of 10^{-4} and weight decay of 10^{-5}. To stabilize training, we use the `ReduceLROnPlateau` learning rate scheduler. This scheduler reduces the learning rate by a factor of 0.5 if the validation loss does not improve for 4 consecutive epochs, ensuring efficient learning without overfitting. We resize all images to 512×512 pixels. Since the default resolution for SAM is 1024×1024, we use bilinear interpolation to resize the positional encoding maps to the new size of 512×512. The LoRA configuration is set with $r = 8$, $\alpha = 16$, and a dropout rate of 0.1. The loss function used is a weighted combination of Dice loss, Focal loss, and Pairwise Consistency with weights set to 0.7, 0.3, and 0.1, respectively. Data augmentation is performed using random rotation and horizontal flipping. The images are normalized to $[-1, 1]$.

For all datasets, we use consistent parameters to ensure fairness in comparison. The threshold for saliency map filtering is set to 0.5, and we consider a total of 5 slices for contextual reasoning.

2. Evaluation Metrics Details

This section elaborates on the evaluation metrics used in our experiments, including their definitions and how special cases were handled during computation.

Region-based metrics. The Dice Similarity Coefficient (DSC) and the Jaccard Index (IoU) measure region-level agreement between the predicted and ground-truth segmentation masks. A score of 1.0 indicates perfect overlap, while 0 denotes no overlap. Both metrics are computed per image and averaged over the test set. Higher values indicate better segmentation quality.

Boundary-based metrics. The 95% Hausdorff Distance (HD95) and Average Surface Distance (ASD) assess boundary alignment between prediction and ground truth. HD95 measures the largest distance between boundaries after excluding the top 5% of outlier points, offering robustness to small errors. Lower values are better. ASD computes the average distance between the predicted and ground-truth surfaces, capturing overall boundary proximity. Again, lower values are preferred.

Handling of empty masks. For cases with empty masks, true negatives are scored as a perfect match, while complete failures receive the worst possible score. Boundary-based metrics such as HD95 and ASD are only computed on samples containing valid segmentation regions.

3. Ablation Study on Contextual Slice Number

We investigate the impact of the contextual slice number n in our Contextual Prompt Distillation (CPD) module, where

n denotes the number of adjacent slices on each side of the current slice. This leads to $2n + 1$ slices as input for contextual reasoning. We vary $n \in \{0, 1, 2, 3\}$ and report the results on the TI dataset.

As shown in Figure 1, performance improves significantly when incorporating contextual information, with the Dice score increasing from 66.95% at $n = 0$ to 73.78% at $n = 1$. Interestingly, while $n = 1$ achieves the highest average performance, the results remain comparable at $n = 2$ (73.58%) and $n = 3$ (71.53%), suggesting that our method is not overly sensitive to the choice of n within a reasonable range.

4. Robustness to Saliency Threshold τ

To validate the stability of our framework, we analyzed its sensitivity to the saliency threshold τ, a key hyperparameter for our prompt distillation. Specifically, we examined the distribution of all predicted saliency values produced by the trained saliency head across the entire TI dataset, as shown in Figure 2.

The histogram reveals a strikingly polarized distribution. A dominant peak appears in the [0.1, 0.2) interval, which alone accounts for 491.3 million pixels—97.88% of all predictions. In contrast, the pixel counts in other bins are several orders of magnitude smaller: for example, only 0.66% of the predictions fall into [0.8, 0.9), and 0.89% into [0.9, 1.0). Intermediate ranges such as [0.3, 0.4), [0.4, 0.5), and [0.5, 0.6) each contribute less than 0.1%. This leads to a virtual "confidence valley" between the dominant low-saliency predictions and the sparse high-saliency signals.

Such extreme skewness indicates that our model is highly confident in distinguishing background from foreground, rarely assigning ambiguous intermediate values. Importantly, any saliency threshold τ chosen within this valley (e.g., from 0.3 to 0.8) will yield nearly identical binarization outcomes and validated prompt sets.

This insensitivity to the exact choice of τ highlights the robustness and hyperparameter stability of our SPD framework, eliminating the need for delicate tuning and simplifying deployment.

5. Comparison with 3D Medical Segmentation Methods

We compare SPD against recent 3D SAM-based methods, SAM2 [2] and MedSAM2 [1], on the TI dataset. As shown in Figure 3, SPD consistently outperforms both SAM2 and MedSAM2 under both Average Dice and Volumetric Dice metrics. Note that Volumetric Dice scores are numerically lower than Average Dice due to severe foreground sparsity across slices.

We additionally report Volumetric Dice scores across all four datasets in Table 1. SPD consistently achieves the highest scores, with clear improvements of +7.34%, +9.08%, +7.30%, and +0.95% over the best prior baseline on TI, Scar, FUMPE, and KiTS, respectively.

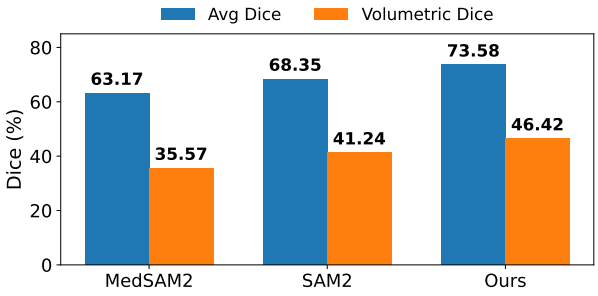

Figure 3. Comparison with recent 3D SAM-based baselines on the Terminal Ileum dataset, evaluated with Average Dice and Volumetric Dice.

Table 1. Volumetric Dice scores. Δ denotes improvement over the best prior baseline.

Dataset	MedSAM	MSA	SAM-Refiner	UNet	nnU-Net	**Ours**	Δ
TI	31.54	34.83	39.08	28.64	14.36	**46.42**	+7.34%
SCAR	32.15	44.60	37.20	47.77	47.66	**56.74**	+9.08%
FUMPE	35.33	46.14	63.09	62.43	66.18	**73.48**	+ 7.30%
KiTS	75.79	91.64	90.20	91.51	93.46	**94.41**	+0.95%

6. Dependence on Stage I

Notably, the saliency head in Stage I is intentionally coarse, achieving only **26.95** Dice when evaluated directly against ground-truth masks. It therefore serves as a weak and imperfect anatomical prior rather than an accurate segmentation module. Despite this coarse Stage I prior, the overall segmentation performance remains strong because the final accuracy is primarily driven by Stage II prompt-guided segmentation. This demonstrates that our framework does not critically depend on the quality of Stage I predictions and confirms the robustness of the two-stage design.

7. Computational Cost

SPD introduces only marginal inference overhead. On the Terminal Ileum dataset using a single NVIDIA RTX 4090, SPD runs at 11.49 FPS, compared to 12.68 FPS for standard SAM. The saliency head is a lightweight decoder, and after consensus prompts are constructed, the final segmentation is performed identically to SAM. In practice, saliency maps and consensus prompts can be computed once per volume and cached, making the effective inference latency comparable to existing SAM-based methods.

References

[1] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei

Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025. 2

[2] 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. 2