SAM2-ProMem: Enhancing Zero-Shot 3D Segmentation with Stochastic Propagation and Memory Search

Yujie Wang[1][0000−0002−2630−6580], Juntao Huang[3], Dazhu Liang[3], Fangzhou Liao[3], Jie Chen[1], and Boan Chen[2(✉)]

[1] Shanghai Jiao Tong University, Shanghai 200240, China
yj.wang@sjtu.edu.cn, dr.jie.chen@aliyun.com
[2] School of AI, Shanghai Jiao Tong University, Shanghai 200240, China
bchen4@sjtu.edu.cn
[3] Shukun Technology, Beijing 100105, China
{huangjt, liangdz, lfz}@shukun.net

Abstract. Although the SAM2 foundational segmentation model excels in natural images, its direct adaptation to 3D medical imaging (e.g., CT/MR) remains underexplored, particularly for zero-shot generalization. We identify two critical barriers when treating medical volumes as pseudo-video sequences: (1) the non-convexity of anatomical structures leading to slice-wise mask discontinuities; (2) difficulty in effectively generalizing the dependencies between long-term and short-term memory. To address these problems, we propose a stochastic connected component propagation strategy for handling mask discontinuities during training, coupled with a dynamic memory window search mechanism during inference. Extensive experiments demonstrate the effectiveness of our method, achieving a 16% Dice score improvement over conventional fine-tuning in the unseen classes of TotalSegmentator dataset. Furthermore, our approach generalizes well across modalities (CT/MR) and lesion types, and it performs comparably to or outperforms previous methods on the ULS23 and CHAOS benchmarks.

Keywords: SAM2 · 3D Medical Image Segmentation · Foundation Model.

1 Introduction

The SAM2 [9], has received significant attention and has been successfully adapted to various domains [7, 8, 11]. It leverages a memory module to integrate historical memories into the segmentation of current frames, enabling effective video segmentation.

This study aims to adapt SAM2, a model originally designed and trained on large-scale video data, to 3D medical image segmentation (primarily CT/MR), with the goal of establishing a universal framework for segmentation in medical volumetric data. Our investigation focuses not only on performance within trained categories, but particularly emphasizes the model's zero-shot capabilities

on unseen anatomical structures. The zero-shot generalization capability holds significant clinical value in addressing the scarcity of annotated medical data. By eliminating the need for task-specific fine-tuning, it can directly segment rare anatomical structures.

Although 3D medical volumes can be processed as pseudo-video sequences, two critical challenges arise. First, unlike natural videos in which object masks are typically temporal continuous, non-convex anatomical structures in medical volumes exhibit mask discontinuities across slices (see Figure 1), causing the model to erroneously bias segmentation towards visually similar regions regardless of spatial separation. Second, the memory dependency mechanism faces a trade-off: a large memory window (e.g., size 6) struggles to distinguish fine-grained structures like adjacent ribs, while a small window (e.g., size 1) degrades performance on complex boundaries (Figure 2). We hypothesize that effectively generalizing long-term and short-term memory dependencies requires large-scale training data. Unfortunately, the scarcity of annotated medical image segmentation data hinders the memory module's generalizability, ultimately reducing zero-shot performance.

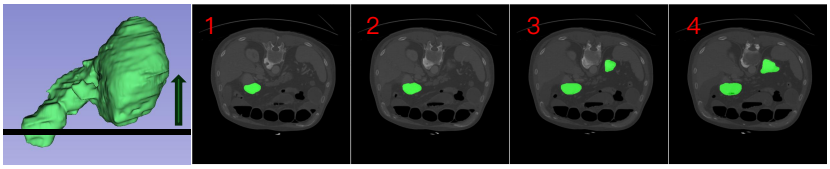

Fig. 1. Left: 3D visualization of a stomach. Slicing upward (in the arrow direction) progressively reveals internal structures. Right: 2D slices with overlaid ground truth masks. Discontinuities occurs in the third slice, where the mask abruptly detaches from the previously tracked region, highlighting segmentation challenges.

Significant research efforts have focused on adapting the SAM2 for medical image segmentation. MedicalSAM2 [12]proposes a self-sorting memory bank that dynamically selects informative embeddings through confidence and dissimilarity metrics. SAM2-Adapter [2] incorporates lightweight adapters in the image encoder, enabling joint fine-tuning with the mask decoder. RFMedSAM2 [10] introduces both an adapter module and a multi-stage automatic prompt refinement framework for medical image segmentation. However, these works do not focus on addressing the aforementioned challenges.

To address these challenges, we propose **SAM2** with stochastic **Pro**pagation and **Mem**ory search (SAM2-ProMem), which introduces two key innovations: (1) a stochastic propagation strategy that enforces spatial coherence during training by retaining only one connected component in cases of mask discontinuities; and (2) an adaptive memory search mechanism that dynamically optimizes the integration of historical context during inference by enumerating multiple

candidate memory window sizes and retaining the top-scoring predictions for each slice.

Experimental results demonstrate a 16% increase in Dice score on unseen classes of the TotalSegmentator dataset[4], compared to direct fine-tuning. Furthermore, comprehensive benchmarking on the ULS23 lesion segmentation dataset and the CHAOS MR dataset shows that our approach performs comparably to or better than state-of-the-art methods.

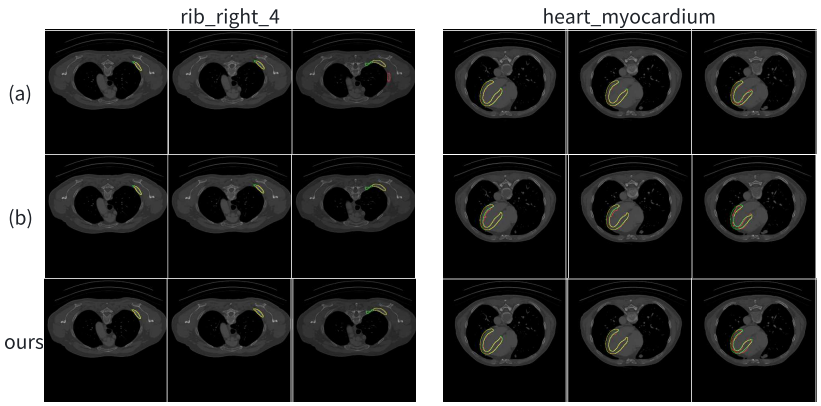

Fig. 2. (a) SAM2 model trained with M3D dataset (memory window size=6) — green contours (ground truth), red (predictions), yellow (overlap). (b) SAM2 trained with M3D dataset (memory window size=1). (c) the proposed SAM2-ProMem method. For rib right 4 segmentation, (a) exhibits duplicate predictions, whereas (b) shows better results. Conversely, heart myocardium segmentation in the right image, (a) yields better segmentation while (b) performs very poorly. (c) demonstrates the robustness of SAM2-ProMem. Best viewed in color.

2 Method

2.1 Stochastic Propagation

The SAM2 architecture, originally developed for video segmentation, processes 2D frames sequentially for mask prediction. To adapt this framework for 3D medical volumes, we reinterpret one anatomical axis (e.g., axial or sagittal) as a pseudo-temporal dimension, thereby preserving SAM2's native slice-wise processing paradigm. However, as demonstrated in Figure 1, conventional subsequence sampling during training may induce mask discontinuities between adjacent slices. This artifact arises from the non-convex morphology of anatomical structures along the pseudo-temporal axis—a phenomenon particularly evident

[4] https://zenodo.org/records/10047292

in complex organ geometries. These discontinuities force the network to predict spatially fragmented regions based on a single preceding region (see slices 2 to 3 in Figure 1), which introduces domain-specific biases that compromise zero-shot generalization capability. To address this challenge, we propose a stochastic propagation strategy comprising three steps:

1. Subsequence Sampling: Sample a subsequence from the 3D masks of a certain class along the pseudo-temporal axis.
2. Discontinuity Detection: Detect discontinuities through connected component analysis. Note that the concept of discontinuity is dependent on the starting slice. For instance, if the starting slice is 1 or 2 (as shown in Figure 4), then slice 3 is considered a discontinuity; however, if the starting slice is 3 or 4, no discontinuity is detected because the upper part in slice 3 or 4 normally disappears when propagating to slice 2.
3. Component Retention: If discontinuities are detected, a single spatially coherent component is stochastically retained during the generation of training instances (see Figure 4).

2.2 Memory Search

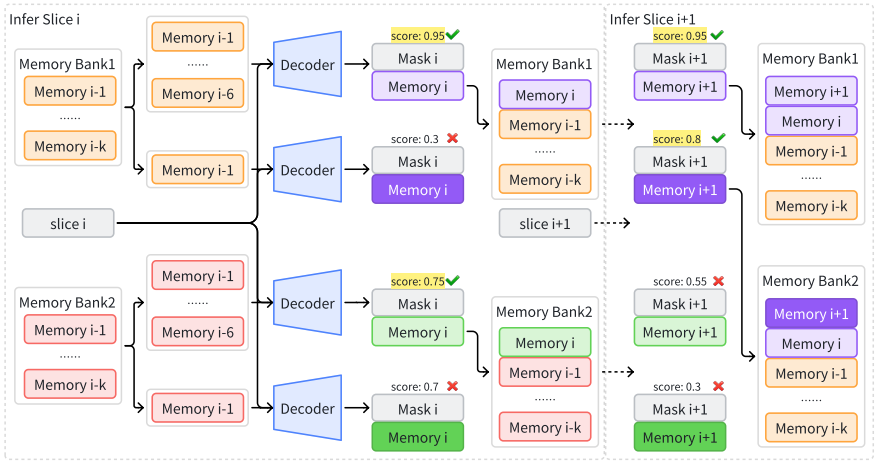

Fig. 3. This figure illustrates the pipeline of the memory window search algorithm. For simplicity, the diagram depicts the scenario with two candidate memory banks. When inferring slice i, the algorithm enumerates combinations of memory banks and the number of historical memories used to compute memory attention. For each combination, it decodes a mask paired with a confidence score. The two highest-scoring predictions are globally selected to update their respective memory banks, while lower-scoring predictions are discarded. The algorithm then proceeds to infer slice i+1 using the updated memory banks. Best viewed in color.

As shown in Figure 3, we employ beam search during inference to dynamically determine the optimal memory window size from a candidate pool comprising N window sizes. The process begins with a single memory bank configuration (not shown in the figure). By iteratively enumerating different window sizes, the algorithm generates multiple predictions, each comprising a segmentation mask and an associated IoU score. The mask with the highest IoU score is selected as the final prediction for the current slice, while the memory states corresponding to the top-k predictions (ranked by IoU) are preserved, forming k distinct memory bank configurations.

When predicting the i-th slice, the algorithm sequentially evaluates all k memory banks. For each configuration, it further explores possible window sizes to compute memory attention, yielding kN candidate results. After globally ranking these candidates, the top-k highest-scoring predictions are retained to update the memory banks for inference on the (i+1)-th slice.

Empirical validation demonstrates that setting k = 2 and N = 2 achieves an optimal trade-off between efficiency and computation. See Figure 5 for two search options. The prediction process continues until the probability of object appearance falls below 0.5, as determined by the configuration with the highest IoU prediction.

Notably, directly applying beam search with the original SAM2 architecture—trained on fixed window sizes—resulted in suboptimal zero-shot performance. To address this limitation, we introduced a novel augmentation strategy during training that stochastically samples memory window sizes and selectively incorporates prompt slice's memory into the memory attention computation. Specifically, the window size is randomly chosen from 1 to 6, and the decision to include the prompt slice memory in the memory attention computation is also determined stochastically.

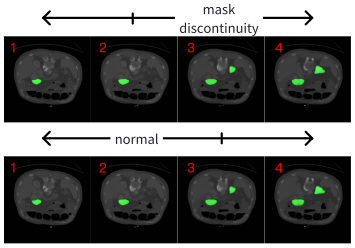

Fig. 4. Illustration of how mask discontinuities depend on the starting slice.

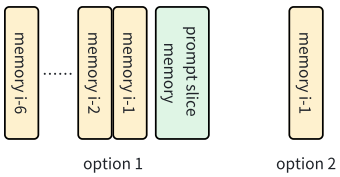

Fig. 5. Illustration of memory window size option.

2.3 Evaluation

During model evaluation, we used a two-stage bidirectional inference protocol. The ground truth mask from the middle slice served as the initial prompt. Forward inference continued until an empty prediction was encountered, after which we cleared the memory banks and performed reverse inference until termination.

3 Experiments

3.1 Dataset

Training Dataset The M3D-Seg dataset [1], a large-scale CT segmentation dataset, contains 25 sub-datasets with 5,772 3D images and 149,196 mask annotations. Each sub-dataset is split into training and testing sets. For training, we used the training splits from 24 sub-datasets, excluding Subset 11 (the TotalSegmentator dataset).

Evaluation Dataset. We conducted two types of evaluations: 1. **In-domain evaluation**. The validation split of the 24 sub-datasets. 2. **Zero-shot evaluation**. Unseen classes from the validation split of M3D Subset 11 were selected due to their diversity and novel class distribution. To further assess zero-shot performance on lesions and MR modality, we also evaluated on the ULS23[4] and CHAOS [5] datasets. If not denoted, average Dice score is reported.

3.2 Implement Details

The SAM2 tiny model[5] is used as the pre-trained model for efficiency. Images are resized so that their longer side is 512 and padded on the shorter side accordingly. A random class is chosen to sample an 8-frame subsequence for training. The model is trained for 28,200 iterations on 8 NVIDIA A10 GPUs (2 sequences per GPU) with the remaining optimization settings identical to SAM2.

3.3 Ablation Study

Impact of Starting Slice Selection. Using the BTCV subset [6], we analyzed how the starting slice position affects performance. As shown in Table 1, selecting a slice near the object's center achieved the highest dice score. Thus, we adopted the central slice and its ground-truth mask as the default prompt.

Role of Stochastic Propagation. Table 2 reveals that naive fine-tuning significantly improves performance. Notably, while disabling stochastic propagation slightly decreased in-domain Dice by 0.4% (Table 2 (b) vs. Table 2 (c)), it reduced zero-shot Dice by 12%, validating stochastic propagation's importance for zero-shot generalization.

Memory Window Search Hyperparameters. For efficiency, we evaluated memory configurations on BTCV [6] (in-domain evaluation) and 30 cases

[5] https://dl.fbaipublicfiles.com/segment_anything_2/072824/sam2_hiera_tiny.pt

Method \ Prompt position	0.1	0.3	0.5	0.7	0.9
fine-tune with SP	0.7632	0.8283	0.8357	0.8343	0.7932

Table 1. Ablation of different prompt slice positions. The header row indicates the normalized position of the prompt slice. For example, 0.5 denotes a prompt slice located at the middle slice of the class ground truth mask.

	(a)SAM2-Tiny(512)	(b) fine-tune w/o SP	(c)fine-tune with SP
In-domain	0.4621	0.7919	0.7879
Zero-shot	0.4649	0.5304	0.6535

Table 2. Ablation of stochastic propagation. SP is short for Stochastic Propagation.

from M3D Subset 11 (zero-shot evaluation). Table 3 and Table 4 demonstrate that including prompt slice memory in memory attention harms zero-shot performance. Since the memory bank size k had minimal impact, we set k=2.

	In-domain	Zero-shot
w/ prompt slice	0.8383	0.6748
w/o prompt slice	0.8335	0.7126

Table 3. Ablation of hyper-parameters of search method. k=2 and N=2.

k	In-domain	Zero-shot
1	0.8320	0.7121
2	0.8335	0.7126
3	0.8324	0.7085

Table 4. Ablation of hyper-parameters of search method, without adding prompt.

Synergy of Memory Augmentation and Search. Table 5 highlights that combining memory augmentation with search yields a 4% zero-shot Dice improvement, whereas using either alone provides marginal gains (0.6%). This underscores their complementary roles.

aug.	search	zero-shot
		0.6535
✓		0.6627
	✓	0.6595
✓	✓	0.7035

Table 5. Ablation of memory augmentation and search mechanism.

	Ground-truth range	Search	Zero-shot
(a)			0.6535
(b)		✓	0.7035
(c)	✓		0.7551
(d)	✓	✓	0.7865

Table 6. Range means limit the prediction in the range of ground truth along the pseudo-temporal axis.

Mask Quality vs. Stop Prediction. From the comparison between (a) and (b) in Table 6, it can be seen that introducing the search method improves the dice score by 5%. However, it is unclear whether this improvement comes

from an enhancement in the 2D mask prediction or from better prediction of the propagation stop position. To clarify this, we fixed the propagation position to the ground-truth range. Compared to (c) and (d), the search method still improved the dice score by 3%, suggesting that it mainly enhances the performance of the 2D mask prediction.

3.4 Comparing with Other Methods

Due to limited resources, we train our model on SAM2-Tiny with a 512-input configuration, while other SAM2-based methods [12, 2, 10] use larger models and higher resolutions. This makes fair comparison challenging, especially given the significant differences in dataset settings. Hence, we primarily compared our work with SegVol [3], which uses a similar data configuration.

In-domain Comparison. Compared to SegVol (trained on M3D), our method shows clear superiority (Table 7). Notably, SegVol was excluded from zero-shot comparison as it used M3D Subset 11 (our zero-shot test set) during training. We use the official SegVol code[6] with box and text prompts to evaluate.

	SegVol	Ours
In-Domain	0.6950	0.7732
Zero-shot	N/A	0.7035

Table 7. In-domain and zero-shot comparison in the M3D dataset.

	SegVol	SAM2-ProMem
Liver	0.8570(0.8319,0.8819)	0.9279(0.9150, 0.9431)
Spleen	0.8009(0.7702,0.8256)	0.9157(0.8908,0.9482)
L-Kidney	0.8004(0.7256,0.8452)	0.9234(0.9061,0.9411)
R-Kidney	0.8146(0.7593,0.8620)	0.9227(0.9018,0.9513)

Table 8. CHAOS MR evaluation in median value of Dice score, i.e. 'Median values (First quartile, Third quartile)'

Zero-shot Comparison. On ULS23 [4], our method achieves competitive median Dice scores (Table 9). For cross-modality evaluation on CHAOS [5] (MRI), we outperform SegVol by a large margin (Table 8), demonstrating strong lesion segmentation and MR adaptation capabilities.

	MedSAM	SAM-MED2D	SAM-MED3D	SegVol	SAM2-ProMem
DeepLesion3D	0.7680	0.3258	0.2386	0.7065	0.7557
BoneLesion	0.6896	0.1947	0.4447	0.6920	0.7416
PancreasLesion	0.6561	0.5548	0.5526	0.7265	0.6235
Average	0.7046	0.3584	0.4120	0.7046	0.7069

Table 9. ULS23 dataset evaluation in median value of Dice score.

[6] https://github.com/BAAI-DCAI/SegVol

4 Conclusion

In this work, we bridge the gap between video-oriented SAM2 and 3D medical image segmentation by addressing two critical challenges: (1) discontinuities in consecutive masks during training, and (2) the balance between long-term and short-term memories. Through our proposed stochastic propagation strategy and the memory window search mechanism, we achieve robust zero-shot generalization across diverse medical targets, including lesions and different modalities.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models (2024)
2. Chen, T., Lu, A., Zhu, L., Ding, C., Yu, C., Ji, D., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more (2024), https://arxiv.org/abs/2408.04579
3. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation (2024)
4. de Grauw, M.J.J., Scholten, E.T., Smit, E.J., Rutten, M.J.C.M., Prokop, M., van Ginneken, B., Hering, A.: The uls23 challenge: a baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography (2024), https://arxiv.org/abs/2406.05231
5. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nürnberger, A., Maier-Hein, K.H., Bozdağı Akar, G., Ünal, G., Dicle, O., Selver, M.A.: Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (Apr 2021). https://doi.org/10.1016/j.media.2020.101950, http://dx.doi.org/10.1016/j.media.2020.101950
6. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge. vol. 5, p. 12. Munich, Germany (2015)
7. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning (2024), https://arxiv.org/abs/2408.07931
8. Qiu, J., Liu, W., Zhang, X., Li, E., Zhang, L., Li, X.: Ded-sam:adapting segment anything model 2 for dual encoder–decoder change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 995–1006 (2025). https://doi.org/10.1109/JSTARS.2024.3490754
9. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024), https://arxiv.org/abs/2408.00714

10. Xie, B., Tang, H., Yan, Y., Agam, G.: Rfmedsam 2: Automatic prompt refinement for enhanced volumetric medical image segmentation with sam 2 (2025), https://arxiv.org/abs/2502.02741
11. Zhou, Y., Sun, G., Li, Y., Benini, L., Konukoglu, E.: When sam2 meets video camouflaged object segmentation: A comprehensive evaluation and adaptation (2024), https://arxiv.org/abs/2409.18653
12. Zhu, J., Qi, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2 (2024)