Leveraging SAM for Generalized Few-Shot 3D Volumetric Segmentation: A Cross-View Propagation Approach

Bocheng Guo[1], Yongyi Su[2], Nanqing Liu[3], Fan Zhang[1], Xulei Yang[4], Xun Xu[4*]
1 University of Electronic Science and Technology China, China
2 South China University of Technology, China
3 School of Information Science and Technology, Yunnan Normal University, Kunming, China
4 Institute for Infocomm Research (I^{2}R), A*STAR, Singapore

***Abstract*—Few-shot segmentation (FSS) enables learning from limited examples, but existing 3D FSS methods often require domain-specific meta-training and may struggle under large structural variations. In this work, we propose Cross-View Propagation with Segment Anything Models (CVP-SAM), a training-free and generalized framework for 3D few-shot volumetric segmentation that repurposes the spatiotemporal segmentation capability of the video foundation models SAM 2 and SAM 3. We reinterpret cross-instance structural differences as temporal deformations in a virtual video, enabling robust label transfer via SAM's built-in tracking and memory mechanisms. Our approach features three key components: (1) Entropy- and Uniformity-Aware Support Optimization, which selects informative support slices by balancing structural uncertainty (via entropy) and spatial coverage; (2) a Virtual Video Bridge that interleaves support and query slices into hybrid sequences for memory-based feature alignment; and (3) Orthogonal Geometric Scouting, a registration-free coarse-to-fine localization strategy using global feature matching across orthogonal views. Experiments on medical and semiconductor 3D volumes show our model-agnostic method achieves state-of-the-art performance with only 5 support slices, demonstrating strong generalization across diverse 3D vision tasks.**

***Index Terms*—Few-Shot Segmentation; 3D Volumetric Segmentation; Segment Anything Model; Virtual Video; Training-Free**

I. Introduction

3D volumetric segmentation supports a wide range of real-world applications, including medical image analysis [1], semiconductor industrial inspection [2], and many others. However, providing ground-truth annotations for 3D volumetric images remains particularly expensive, as each volume contains a large number of slices and annotation is typically performed slice-by-slice. Consequently, exhaustively labeling all slices becomes highly inefficient for large-scale 3D segmentation tasks. This high annotation cost has motivated data-efficient approaches, particularly few-shot segmentation, which aims to leverage only a small number of labeled slices.

* Correspondence to Xun Xu <xu_xun@a-star.edu.sg>.
This research work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

Existing few-shot segmentation methods formulate the problem from different perspectives, including feature refinement [3], morphological correlation modeling [4], and leveraging 2D segmentation priors [5], [6]. Despite the notable progress, these approaches require model training or fine-tuning, which hinders fast and reliable deployment due to additional computational overhead, especially when applied to large 3D volumes.

The emergence of powerful 2D segmentation foundation models, such as the Segment Anything Model (SAM) [7], has stimulated increasing interest in adapting these general-purpose models for few-shot 3D segmentation. For instance, MSFSeg reduces 3D segmentation into slice-level segmentation, cross-slice correlation, and fusion [4], while RadSAM propagates prompts from a single slice to the entire volume to improve label efficiency [8]. These methods substantially reduce the need for dense voxel-level annotations by allowing annotators to label only a few slices or even provide a single prompt. However, many anatomical structures exhibit complex 3D shapes, such as branching vessels, multi-slice tumors, or irregular lesions, that are difficult to capture using slice-based modeling alone. When annotated support slices are unrepresentative or fail to capture structural variability, segmentation performance may degrade due to insufficient volumetric context.

To fully address these issues, we propose a training-free and fully 3D-aware framework that avoids model fine-tuning altogether. Our approach is inspired by the idea of formulating 3D volumetric segmentation as a video object segmentation task. Recent advances such as S2VNet [9], SliceProp [10], and iSegFormer [11] demonstrate that space–time memory, correspondence propagation, and temporal feature aggregation are effective for maintaining consistent segmentation across sequential data. Motivated by these insights, we reinterpret a 3D volume as a temporal sequence of slices, enabling the use of the video segmentation foundation models SAM 2 [12] and SAM 3 [13] as unified space–time segmentation backbones. This perspective provides robust cross-slice propagation, reduces annotation effort, and improves 3D structural coherence without requiring computationally heavy 3D architectures.

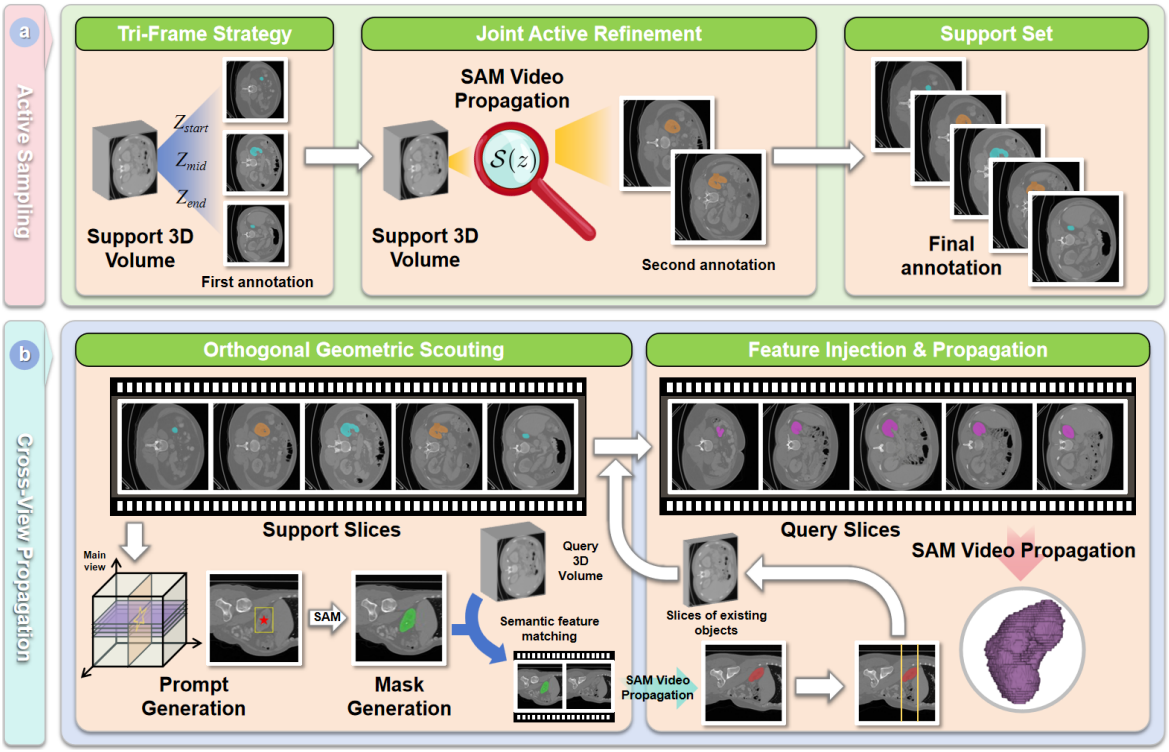

Fig. 1. Overview of the proposed Cross-View Propagation Framework. The pipeline consists of three sequential stages: 1) Support Set Construction (top row); 2) Orthogonal Geometric Scouting for query localization (bottom left); and 3) Feature Injection via a "Virtual Video Bridge" (bottom right).

We further enhance the effectiveness of the framework from complementary angles. First, recognizing that support slices are pivotal in few-shot segmentation, we design an effective guideline for support-slice selection following two intuitive principles: diversity and model uncertainty [14]. Second, we develop a robust strategy to propagate segmentation masks from the annotated support slices to the query slices. A key component is the Cross-View Propagation mechanism, which identifies matching slice pairs between support and query volumes and transfers labels/masks accordingly. To overcome localization challenges in unaligned or heterogeneous volumetric data, we introduce Orthogonal Geometric Scouting, a coarse-to-fine localization strategy leveraging global semantic feature matching from orthogonal projections (e.g., sagittal and coronal views) without explicit 3D registration. To facilitate reliable label transfer, we construct a Virtual Video Bridge that forms hybrid support–query slice sequences, enabling precise propagation through the SAM's memory bank. We refer to the proposed method **CVP-SAM** to respect the core idea of propagation across viewpoints. Our main contributions are summarized as follows:

- We propose a model-agnostic, training-free framework that leverages SAM 2 and SAM 3 for 3D volumetric segmentation, combining temporal tracking with cross-instance structural transfer.
- We introduce Orthogonal Geometric Scouting, a coarse-to-fine localization strategy that avoids heavy 3D registration, along with an uncertainty-aware support-set optimization.
- We validate the framework's versatility by achieving state-of-the-art results with either SAM 2 or SAM 3, highlighting its potential as a universal 3D analysis tool.

II. Method

A. Overview

We leverage the state-of-the-art video object segmentation foundation models, i.e. SAM 2 [12] and SAM 3 [13], for few-shot 3D volumetric segmentation. As illustrated in Fig. 1, our framework consists of three stages. First, we construct the support set by selecting informative slices for labeling. Second, we introduce orthogonal geometric scouting to localize the object of interest in the support set. Finally, we propagate the masks to query slices to achieve few-shot segmentation.

B. Support Set Construction

To ensure the labeled slices are both anatomically representative and spatially comprehensive, we construct the support set in two stages. First, we initialize the set S_{init} using a fixed selection strategy, Tri-Frame Strategy, which simply selects the start ($z_{\min}^{\text{supp}}$), middle, and end ($z_{\max}^{\text{supp}}$) slices of the organ's spatial extent from the ground truth.

Subsequently, we employ an Entropy-Spatial Joint Active Refinement to augment S_{init} with N_{supp} slices, balancing uncertainty and coverage. We first perform self-propagation using S_{init} to estimate the prediction ambiguity of unlabeled slices via Shannon entropy:

$$H(z) = -\sum_{c\in\{0,1\}} p_c(z)\log p_c(z) \tag{1}$$

where $p_c(z)$ denotes the SFM's predicted probability of class c (foreground/organ = 1, background = 0) at slice z. A higher $H(z)$ indicates greater anatomical ambiguity, e.g. complex branching vessels, blurred boundaries, and thus higher value for inclusion in the support set.

To prevent spatial bias, we identify "spatial gaps" by encouraging selecting equidistant slices within the volume. For each gap centered at p with a span of d_{span}, we select the optimal slice z^* by maximizing a joint score:

$$\mathcal{S}(z) = \omega \cdot \frac{H(z)}{\max(H)} + (1-\omega) \cdot \max\left(0, 1 - \frac{|z-p|}{d_{\text{span}}}\right) \quad (2)$$

where the first term prioritizes anatomical ambiguity (normalized entropy) and the second term enforces spatial proximity to the gap.

The selection process iterates until enough slices are selected (or no suitable slices remain), forming the optimized support set $S_{\text{opt}} = S_{\text{init}} \cup \{z_1^*, z_2^*, ..., z_{N_{\text{supp}}}^*\}$ (typically 5 slices total). By prioritizing slices that fill spatial gaps while retaining high entropy, S_{opt} captures both ambiguous anatomical features and global volumetric coverage.

C. Orthogonal Geometric Scouting

Localizing the object of interest, e.g. query organ or component, in an unaligned volume is a major challenge in few-shot segmentation. We propose a registration-free localization strategy using orthogonal projections, which aligns subjects based on deep semantic topology rather than raw pixel intensity.

Prompt Generation: Utilizing the full 3D geometric profile of the support object derived from all labeled slices in S_{opt}, we first aggregate the bounding boxes and centroids of every slice in S_{opt} to compute the global 3D bounding box (B_{supp}^{3D}) and volumetric centroid (C_{supp}^{3D}) of the support organ. We then project this holistic 3D geometric cue, i.e. B_{supp}^{3D} and C_{supp}^{3D} onto an *orthogonal view*, e.g. sagittal plane for z-axis propagation, to generate a semantic "scout prompt" that encodes the organ's complete topological signature. Both the centroid point and bounding box are used as visual prompt for SAM to segment the object of interest.

Semantic Feature Matching: We utilize the image encoder Φ from the Segment Anything Model (SAM) to extract global semantic embeddings and identify the matching slice in the query volume. Specifically, for the support scout slice in an orthogonal view, we compute $\Phi(I_{\text{supp}}^{\text{scout}})$. To select a unique slice within this orthogonal view, we use the centroid of C_{supp}^{3D} to determine the position along the y-axis, as shown in Fig. 1. For the query volume, we exhaustively evaluate all slices along the orthogonal axis, denoted as $\{\Phi(I_{\text{qry}}^k)\}$. Each slice is compared against the support scout slice. The best-matching query slice $\hat{k}$ is identified by maximizing cosine similarity:

$$\hat{k} = \arg\max_k \frac{\Phi(I_{\text{supp}}^{\text{scout}}) \cdot \Phi(I_{\text{qry}}^k)}{\|\Phi(I_{\text{supp}}^{\text{scout}})\| \cdot \|\Phi(I_{\text{qry}}^k)\|} \quad (3)$$

Crucially, the Perception Encoder is trained via large-scale contrastive vision-language alignment. This ensures that the global semantic embeddings of the query organ remain invariant across different subjects, making our scouting strategy significantly more reliable against inter-subject intensity and shape variations.

Localization: We construct a two-frame "virtual video" comprising the support scout slice (overlaid with the projected B_{supp}^{3D} and C_{supp}^{3D} as geometric prompts, plus the ground-truth mask of the support organ) and the matched query slice $\hat{k}$. The holistic support geometric prompt guides the SAM to propagate the mask to the query slice, from which we parse the 2D bounding box of the query organ in $\hat{k}$. By extending this 2D box along the query's main axis (z-axis) and validating with semantic consistency checks, we finally determine the precise 3D spatial extent $[z_{\text{min}}^{\text{qry}}, z_{\text{max}}^{\text{qry}}]$ of the query organ, eliminating the need for explicit 3D registration.

D. Virtual Video Bridge for Feature Injection

This stage performs the core cross-subject transfer by mapping the optimized support set S_{opt} to the query domain and leveraging the SAM's memory mechanisms to bridge anatomical variations. First, we linearly interpolate the positions of slices in S_{opt} within their respective valid anatomical ranges, specifically mapping from the support's organ-specific z-range $[z_{\text{min}}^{\text{supp}}, z_{\text{max}}^{\text{supp}}]$ (i.e., the spatial extent of the support organ) to the query's organ-specific z-range $[z_{\text{min}}^{\text{qry}}, z_{\text{max}}^{\text{qry}}]$ (the localized spatial extent of the query organ). This generates corresponding slice pairs $(I_{\text{supp}}^i, I_{\text{qry}}^i)$ for each $i \in S_{\text{opt}}$, where I_{supp}^i denotes the i-th slice from the support's organ range (with its ground-truth mask M_{supp}^i) and I_{qry}^i denotes the spatially aligned slice within the query's organ range, ensuring the mapping is constrained to the relevant anatomical regions rather than the global volumetric space.

E. Volumetric Propagation within Query

Through the Virtual Video Bridge, we obtain a sparse set of high-quality pseudo-labels ($|\mathcal{M}_{\text{pseudo}}| = N_{\text{total}} \approx 5$ slices) in the query volume. To generate a dense volumetric segmentation, we perform propagation within query volume by treating the entire query volume V_{qry} as a continuous video sequence along the z-axis. The pseudo-labels $\mathcal{M}_{\text{pseudo}} = \{M_{\text{qry}}^i\}_{i \in S_{\text{opt}}}$ are further injected into the SAM's memory bank as Conditioning Frames, anchoring the model to the query domain's anatomical landmarks and ensuring consistency. The SAM then performs bidirectional propagation along the z-axis. For any unlabeled slice I_t (where $t \in [z_{\text{min}}^{\text{qry}}, z_{\text{max}}^{\text{qry}}]$), the segmentation mask is inferred by attending to accumulated memory features from both past and future conditioning frames:

$$M_t = \text{Attention}(I_t, \mathcal{M}_{\text{past}} \cup \mathcal{M}_{\text{future}}) \quad (4)$$

where $\mathcal{M}_{\text{past}}$ denotes memory embeddings from conditioning frames before slice t, and $\mathcal{M}_{\text{future}}$ denotes embeddings from frames after t. Leveraging the SAM's advanced tracking capabilities, particularly SAM 3's robustness to object disappearance, reappearance, and topological changes, this step

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS ON THREE VOLUMETRIC DATASETS. **THE ROW "TRAINING DATA"** INDICATES THE NUMBER OF LABELED VOLUMES USED FOR TRAINING THE FEW-SHOT MODELS (META-TRAINING) ON EACH DATASET.

Method	**Backbone**	**Training**	**Synapse (Dice %)** *(Train: 5 Vols)*				**MSD (Dice %)** *(Train: 3 Vols)*		**XRM (Dice %)** *(Train: 5 Vols)*
			Spleen	Liver	Kidney	Pancreas	Heart	Hippocampus	Mean
Registration-based Methods									
ANTs [15]	—	×	81.73	85.69	80.58	24.37	78.44	72.37	49.41
Few-Shot Methods (Meta-training Required)									
SAM2-Adapter [16]	SAM2	✓	86.26	91.69	82.57	46.79	81.51	74.38	57.32
SAM3-Adapter [16]	SAM3	✓	89.41	93.93	84.24	61.67	83.88	80.77	64.43
H-SAM [17]	SAM	✓	90.87	93.69	82.94	50.92	80.10	77.75	63.12
PG-SAM [18]	SAM	✓	88.43	92.27	83.26	46.43	79.21	71.99	—
General Foundation Models (Training-Free)									
PerSAM [**?**]	SAM3	×	68.80	88.06	75.19	34.58	71.44	76.53	53.57
Matcher [19]	DINOv3 & SAM3	×	72.42	89.69	66.74	26.76	78.38	76.77	51.21
CVP-SAM (Ours)	SAM 2	×	91.43	92.59	85.04	56.48	82.06	78.23	58.27
CVP-SAM (Ours)	SAM 3	×	**93.50**	**95.37**	**92.03**	**65.62**	**85.41**	**81.73**	**72.13**

effectively handles complex anatomical structures, e.g. discontinuous vessels, nested organs, and ensures z-axis consistency across the entire query volume.

Each pair is treated as a two-frame virtual video, where the SAM propagates the support ground-truth mask to the query slice:

$$M^i_{\text{qry}} = \mathcal{T}(I^i_{\text{supp}}, M^i_{\text{supp}}, I^i_{\text{qry}}; \Theta_{\text{SAM}}) \tag{5}$$

where $\mathcal{T}$ denotes the SAM's mask propagation function parameterized by Θ_{SAM}.

The Memory Bank mechanism plays a pivotal role in this transfer: the support ground-truth mask M^i_{supp} is encoded into compact, semantically invariant memory embeddings. For SAM 3, this leverages its Spatial Memory mechanism, empowered by the semantic consistency of the Perception Encoder, to be robust to local occlusions and anatomical deformations. Even when the query organ exhibits significant shape divergence from the support, the semantically aligned features ensure stable memory retrieval. The SAM reconstructs the query mask M^i_{qry} by querying these support memory embeddings in the latent space, effectively bypassing the domain gap caused by non-rigid anatomical variations between subjects.

III. EXPERIMENTS

A. Experimental Setup

Datasets: We evaluate our framework on three datasets. Medical Segmentation Decathlon **(MSD)** [20] provides diverse 3D biomedical tasks, testing robustness to anatomical variability, different fields of view, and imaging modalities. Synapse Multi-Organ CT **(Synapse)** [21] contains challenging abdominal CT scans with low contrast and complex organ layouts, assessing boundary delineation in crowded anatomical regions. To test cross-domain generalization, we introduce a proprietary Industrial HBM X-Ray Microscopy **(XRM)** dataset [22], featuring micro-scale scans of HBM packages with multi-material structures, internal defects, and strong metallic artifacts, representing a distribution other than medical images. We use the Memory Die subset for evaluations.

TABLE II

ABLATION ON THE SYNAPSE DATASET FOR THE SCOUTING MODULE AND ALTERNATIVE ACTIVE STRATEGIES.

Backbone	**Scout**	**Active Strategy**	**Dice % ↑**
SAM 2	×	Random selection	47.81
SAM 2	✓	Random selection	59.24
SAM 2	✓	K-center	60.36
SAM 2	✓	Entropy Only (Uncertainty)	61.73
SAM 2	✓	Spatial Only (Uniformity)	64.39
SAM 2	✓	**Ours (Joint Entropy-Spatial)**	**66.17**
SAM 3	✓	**Ours (Joint Entropy-Spatial)**	**71.24**

Evaluation Metrics: We use Dice to measure region overlap, with higher values indicating better performance.

Implementation Details: Our framework is entirely training-free, using pre-trained SAM 2 and SAM 3 checkpoints as backbones without fine-tuning, with inference performed on a single NVIDIA 3090 GPU. A detailed inference efficiency analysis and comparison with baseline paradigms are provided in the Supplementary Material. For the Tri-Frame Initialization, slices are selected at the start, middle, and end of the support geometry. During the Uncertainty-Aware Refinement stage, 2 additional frames are added based on entropy ranking, resulting in a total of 5 support slices per volume. The orthogonal scouting module uses the sagittal view by default.

B. Comparison with State-of-the-Art Methods

We benchmark our proposed framework, **CVP-SAM**, against three distinct categories of segmentation approaches to validate its effectiveness: Registration-based methods, where we utilize widely adopted tools such as ANTS [15], performing segmentation by aligning a labeled support atlas to

the query volume via deformable registration; Meta-learning-based FSS methods, including specialized few-shot segmentation models such as SAM-Adapter [16], H-SAM [17], and PG-SAM [18], which typically require episodic meta-training on base classes to learn transferable prototypes; and Training-free Foundation Model adaptations, encompassing recent strategies like PerSAM [**?**] and Matcher [19]. Crucially, to ensure a fair comparison and establish rigorous baselines, we re-implemented these methods by replacing their original backbones with the state-of-the-art SAM 3, ensuring that all foundation-model-based baselines benefit from the most advanced semantic representations available, thereby isolating the contribution of our proposed cross-view propagation strategy. Table I summarizes the quantitative results on our datasets. Our framework, using SAM 2 and SAM 3 as the backbone, outperforms registration-based methods by a significant margin, primarily because our feature-space alignment via the *Virtual Video Bridge* handles non-rigid deformations better than pixel-space registration. Furthermore, we achieve competitive or superior performance compared to meta-learning-based FSS methods. Visual comparisons are presented in 2. As observed, our method generates masks with superior boundary adherence and topological consistency compared to H-SAM and SAM3-Adapter, particularly in regions with low contrast or complex shapes. While baseline methods often fail to capture fine details or exhibit segmentation leakage, our approach effectively preserves the structural integrity of the organ.

C. Ablation Study

To validate the effectiveness of each component in our framework, we conducted an extensive ablation study on the Synapse dataset, with results reported in Table II. Comparing the blind transfer baseline (Row 1) with the scouting-enabled setting (Row 2), we observe a substantial performance drop of 11.4% when Orthogonal Geometric Scouting is removed, indicating that coarse-to-fine localization plays a critical role in handling unaligned volumetric data and aligning semantic topology prior to feature injection. We further examine different support-slice sampling criteria (Rows 2–6). K-means clustering (Row 3) yields a slight improvement over random selection (Row 2) by increasing feature diversity, but remains inferior to spatially explicit strategies. Spatial uniformity (Row 5, 64.39) outperforms entropy-only sampling (Row 4, 61.73), suggesting that global volumetric coverage is more influential than uncertainty-based sampling alone. The joint strategy (Row 6) achieves the best performance, exceeding the strongest single criterion (spatial uniformity) by 1.8% and random selection by 7.0%, demonstrating that jointly balancing boundary uncertainty and spatial coverage leads to more robust adaptation. Finally, replacing the backbone with SAM 3 (Row 7) further improves performance, achieving a DSC of 71.24. This gain can be attributed to the enhanced spatial memory mechanism of SAM 3, which mitigates mask drift in long volumetric sequences by retaining occlusion-aware features along the z-axis.

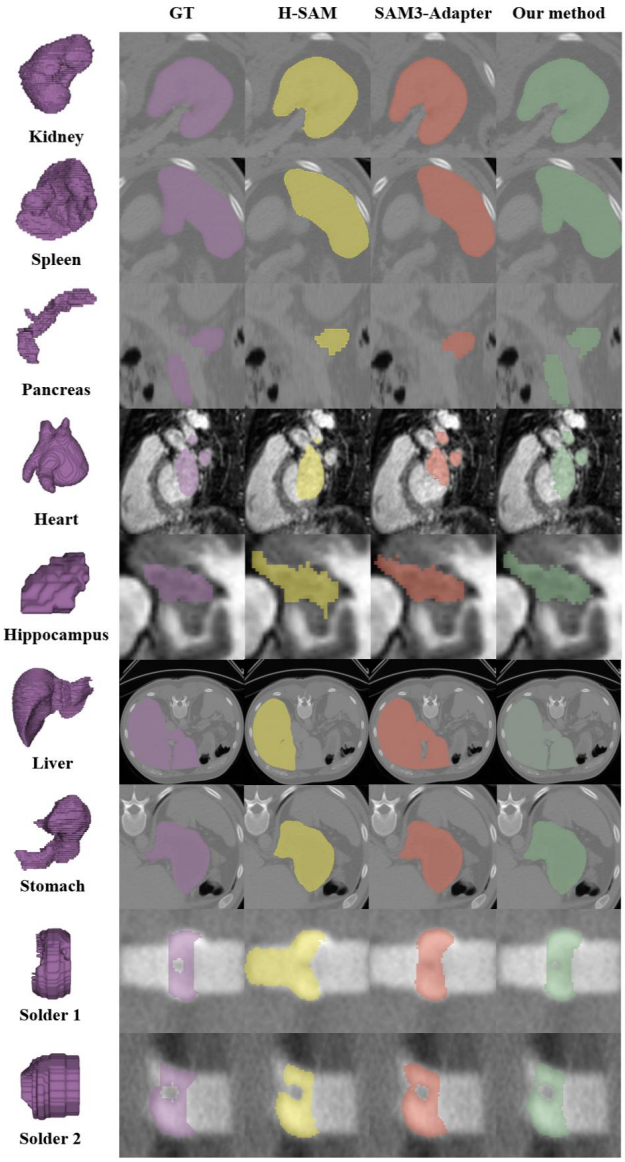

Fig. 2. Qualitative comparisons on representative organs and semiconductor components against state-of-the-art methods.

D. Qualitative Study

We present qualitative comparisons between our method and state-of-the-art SAM-based approaches. As shown in Fig. 2, our method consistently outperforms H-SAM and SAM3-Adapter. Notably, it excels at segmenting elongated anatomical structures that often appear fragmented across 2D slices, such as the arteries in the heart and the two ends of the pancreas, producing more connected and anatomically plausible masks. Moreover, our results exhibit smoother boundaries than those of existing methods, highlighting a key advantage of formulating volumetric segmentation as a video object segmentation task: temporal (inter-slice) consistency naturally encourages spatial smoothness. Additional visualizations, particularly demonstrating the robustness of the Orthogonal Geometric Scouting module, are available in the Supplementary Material.

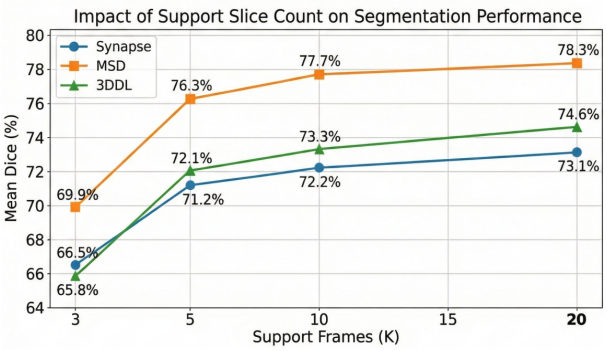

Fig. 3. Impact of support set size (K) on segmentation performance across three datasets. The curve illustrates that increasing K from 3 (geometric initialization only) to 5 (with active refinement) yields the most significant performance gain. Further increasing K to 10 or 20 results in diminishing marginal returns, validating the efficiency of our default setting ($K = 5$).

E. Impact of Support Set Size

We investigated the sensitivity of our framework to the number of annotated support slices, with performance on three datasets reported in Figure 3 for varying support set sizes ($K = \{3, 5, 10, 20\}$). As shown, increasing the support set size from 3 to 5 yields a significant performance boost (e.g., +6.4% on MSD), validating the effectiveness of our *Uncertainty-Aware Active Refinement*, which successfully identifies and corrects the most ambiguous regions missed by geometric sampling. However, further increasing K to 10 or 20 results in diminishing marginal gains, confirming that our framework is highly data-efficient, capturing the majority of anatomical variance with just 5 carefully selected slices.

IV. Conclusion

This paper presents a general, training-free framework for few-shot 3D volumetric segmentation by modeling cross-instance anatomical variation as a virtual spatiotemporal deformation. The pipeline integrates entropy–spatial joint support set optimization, orthogonal-view semantic scouting, and virtual video–based feature transfer via SAM. This formulation allows pre-trained video segmentation models, such as SAM 2 and SAM 3, to propagate volumetric labels without task-specific fine-tuning or meta-training. Evaluations on multiple medical and industrial datasets show consistent improvements over training-free baselines and performance comparable to heavily supervised methods, suggesting that spatiotemporal memory mechanisms in large video models can effectively replace conventional 3D alignment or meta-learned metric spaces for few-shot volumetric analysis.

References

[1] S. Niyas, S. Pawan, M. A. Kumar, and J. Rajan, "Medical image segmentation with 3d convolutional neural networks: A survey," *Neurocomputing*, 2022.

[2] R. S. Pahwa, R. Chang, W. Jie, X. Xun, O. Z. Min, F. C. Sheng, C. S. Choong, and V. S. Rao, "Automated detection and segmentation of hbms in 3d x-ray images using semi-supervised deep learning," in *IEEE Electronic Components and Technology Conference*, 2022.

[3] S. Hansen, S. Gautam, S. A. Salahuddin, M. Kampffmeyer, and R. Jenssen, "Adnet++: A few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement," *Medical Image Analysis*, 2023.

[4] M. Zheng, B. Planche, Z. Gao, T. Chen, R. J. Radke, and Z. Wu, "Few-shot 3d volumetric segmentation with multi-surrogate fusion," in *MICCAI*, 2024.

[5] H. Ding, C. Sun, H. Tang, D. Cai, and Y. Yan, "Few-shot medical image segmentation with cycle-resemblance attention," in *WACV*, 2023.

[6] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "'squeeze & excite'guided few-shot segmentation of volumetric images," *Medical Image Analysis*, 2020.

[7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023.

[8] J. Khlaut, E. Ferreres, D. Tordjman, H. Philippe, T. Boeken, P. Manceron, and C. Dancette, "Radsam: Segmenting 3d radiological images with a 2d promptable model," in *MICCAI*, 2025.

[9] Y. Ding, L. Li, W. Wang, and Y. Yang, "Clustering propagation for universal medical image segmentation," in *CVPR*, 2024.

[10] X. Xu, W. Lu, J. Lei, P. Qiu, H.-B. Shen, and Y. Yang, "Sliceprop: A slice-wise bidirectional propagation model for interactive 3d medical image segmentation," in *IEEE International Conference on Medical Artificial Intelligence*, 2023.

[11] Q. Liu, Z. Xu, Y. Jiao, and M. Niethammer, "isegformer: interactive segmentation via transformers with application to 3d knee mr images," in *MICCAI*, 2022.

[12] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," in *ICLR*, 2025.

[13] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer, "Sam 3: Segment anything with concepts," in *ICLR*, 2026.

[14] L. Cai, X. Xu, L. Zhang, and C.-S. Foo, "Exploring spatial diversity for region-based active learning," *IEEE Transactions on Image Processing*, 2021.

[15] B. B. Avants, N. Tustison, G. Song *et al.*, "Advanced normalization tools (ants)," *Insight j*, 2009.

[16] T. Chen, R. Cao, X. Yu, L. Zhu, C. Ding, D. Ji, C. Chen, Q. Zhu, C. Xu, P. Mao, and Y. Zang, "Sam3-adapter: Efficient adaptation of segment anything 3 for camouflage object segmentation, shadow detection, and medical image segmentation," *arXiv preprint arXiv:2511.19425*, 2025.

[17] Z. Cheng, Q. Wei, H. Zhu, Y. Wang, L. Qu, W. Shao, and Y. Zhou, "Unleashing the potential of SAM for medical adaptation via hierarchical decoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[18] Y. Zhong, Z. Luo, C. Liu, F. Tang, Z. Peng, M. Hu, Y. Hu, J. Su, Z. Ge, and I. Razzak, "PG-SAM: Prior-guided SAM with medical for multi-organ segmentation," *arXiv preprint arXiv:2503.18227*, 2025.

[19] Y. Liu, M. Zhu, H. Li, H. Chen, X. Wang, and C. Shen, "Matcher: Segment anything with one shot using all-purpose feature matching," in *ICCV*, 2024.

[20] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *Nature Communications*, vol. 10, no. 1, p. 2635, 2019.

[21] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[22] R. S. Pahwa, M. T. Lay Nwe, R. Chang, O. Z. Min, W. Jie, S. Gopalakrishnan, D. H. Soon Wee, R. Qin, V. S. Rao, H. Dai, J. T. Neumann, R. Pichumani, and T. Gregorich, "Automated attribute measurements of buried package features in 3d x-ray images using deep learning," in *IEEE Electronic Components and Technology Conference*, 2021.