This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Bio2Vol: Adapting 2D Biomedical Foundation Models for Volumetric Medical Image Segmentation

Jiaxin Zhuang[1], Linshan Wu[1], Xuefeng Ni[1], Xi Wang[1], Liansheng Wang[2], and Hao Chen[1] ✉

[1] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China.
{jzhuangad,linshan.wu,vancywangxi,jhc}@cse.ust.hk, nixuefeng@hnu.edu.cn
[2] School of Information Science and Engineering, Xiamen University, Xiamen, China.
lswang@xmu.edu.cn

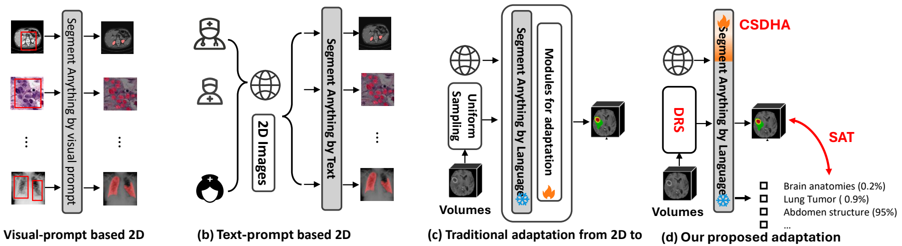
Abstract. 2D biomedical foundation models (FM) have demonstrated remarkable capabilities in 2D medical image segmentation across various modalities, with text-prompted approaches offering scalable analysis that facilitate integration with LLMs and clinical application. Adapting these models for 3D medical image segmentation can leverage their rich visual features while enabling text-prompted volumetric image segmentation. However, efficient adaptation poses significant challenges due to the substantial disparity between 2D and 3D medical images and the necessity to establish text-volume alignment. To address these limitations, we propose **Bio2Vol**, a novel adaptation framework that enables text-prompted 2D biomedical FMs to effectively handle volumetric data. Specifically, (1) To bridge the dimensional disparity, we propose a Dual-Rate Sampling strategy (DRS) that processes inter slices within a volume at both sparse and dense intervals, capturing global contexts and local details; (2) To enhance volumetric feature representation, a Cross-slice Dual-head Attention (CSDHA) is built upon the intra-slice features by repurposing existing pre-trained attention modules for parameter-efficient inter-slice information fusion; and (3) To establish text-volume understanding, a Semantic Text-Visual Alignment loss (SAT) is used to extend the existing 2D text-visual alignment to the volumetric domain. Using BiomedParse as a demonstration case, extensive evaluation across 11 medical datasets across diverse anatomical regions and modalities shows that Bio2Vol significantly improves 3D medical image segmentation performance, enhancing DSC by 4.72% on Amos22 dataset with substantial improvements across MSD tasks. Code will be available https://github.com/JiaxinZhuang/Bio2Vol.

Keywords: Adaptation · Foundation Model · 3D Medical Images.

1 Introduction

Medical image segmentation is fundamental to biomedical discovery, supporting clinical diagnosis, surgical planning, and disease monitoring [21, 36, 4, 20].

Foundation models have catalyzed a paradigm shift in medical image analysis, advancing automation and accuracy [26, 5, 23, 15, 32, 7]. However, specialized FMs remain limited by modality-specific training requirements [26, 5, 23], restricting their universal applicability [15, 32, 18]. Recent innovations have expanded multi-modality support and improved interaction mechanisms [32, 15, 17] for 2D medical images. While SAM-based methods [15] work across various 2D medical image types, they require manual *visual prompts*. In contrast, BiomedParse [32] enables fully automatic segmentation through *text prompts* alone and supporting nine different 2D medical imaging modalities. This transition from visual to text-based prompting eliminates the need for manual inputs, enhancing clinical deployability where efficiency and reproducibility are essential [32, 28, 33]. By pre-training over 6 million 2D image-mask-description triples across nine modalities with segmentation, detection, and recognition tasks, BiomedParse [32] yields robust biomedical knowledge and language understanding capabilities that outperform general-purpose models adapted to medical domains. Despite its effectiveness with 2D medical images, these biomedical FMs exhibit limitations when applied to 3D volumetric datasets, as they cannot capture inter-slice contextual dependencies in the volumetric data and its alignment with text. This constraint is particularly problematic given the prevalence of volumetric data in clinical diagnostic imaging [19, 31, 3, 28, 25, 10, 16]. Although recent 3D medical image FMs [8, 2, 18, 25, 38] have improved performance on volumetric medical images, they are typically restricted to 3D applications, losing the versatility of working with 2D medical image modalities [3, 25]. Furthermore, these models require specialized 3D encoders that cannot leverage the rich multi-modal knowledge learned from the vast amount of 2D medical imaging data [5, 19, 31].



(a) Visual-prompt based 2D biomedical segmentation foundation model (e.g., SAM, etc.)

(b) Text-prompt based 2D biomedical segmentation foundation model, (e.g., BiomedParse, etc.)

(c) Traditional adaptation from 2D to 3D medical images for BiomedParse (e.g., ST-Adapter, AIM)

(d) Our proposed adaptation method for BiomedParse (Bio2Vol)

Fig. 1: Comparison of prompt-guided 2D biomedical foundation models for medical image segmentation and text-prompted adaptation to 3D medical image.

Adapting text-prompted 2D FMs to 3D domains presents three critical challenges: (1) Volumetric Context - understanding 3D context by capturing inter-slice relationships and anatomical continuity while maintaining powerful pre-trained 2D representations; (2) Knowledge Preservation - establishing text knowledge alignment with 3D volume context beyond original pre-trained 2D representations; and (3) Computational Efficiency - balancing performance with computational costs. As shown in Fig.1, existing approaches typically adopt parameter-efficient methods by freezing the pre-trained 2D biomedical FM and introducing additional adapters such as memory modules to information of the previous slices [25], a spatial-temporal operator by depth-wise 3D convolution layers between bottlenecks[19, 18, 10] or progressively joint-adaptations for each two dimensions respectively [31]. However, these methods face significant limitations. They only finetune the model with uniformly sampled slices from the volume, failing to comprehensively capture both global context and local details needed for proper volumetric understanding. Furthermore, these adaptation approaches don't extend text-image understanding to text-volume understanding, preventing effective knowledge preservation between text prompts and volumetric features. Additionally, they introduce extra parameters to adapt the 2D biomedical FM, sacrificing computational efficiency.

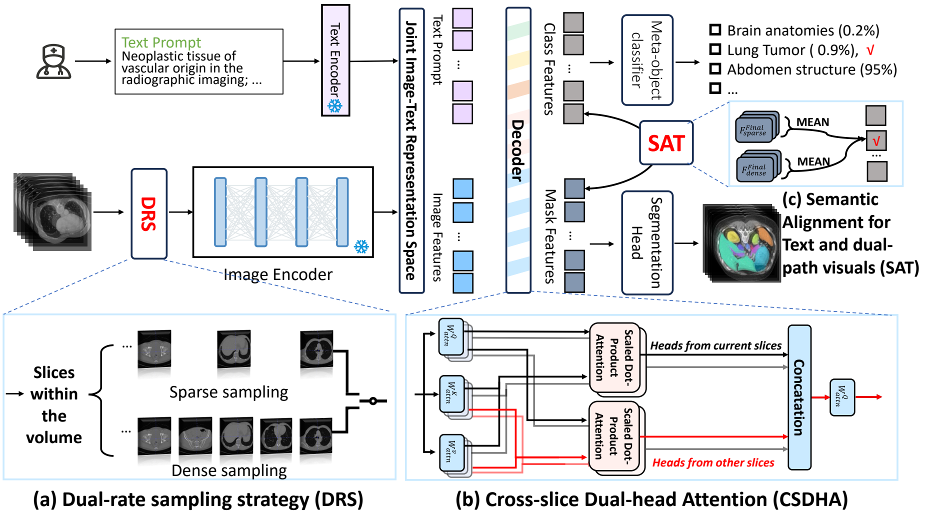
This paper introduces Bio2Vol, a novel adaptation framework that extends text-prompted 2D biomedical FMs such as BiomedParse [32] to volumetric medical image segmentation. As illustrated in Fig. 1(d), our approach bridges the 2D-3D gap through three synergistic components: (1) DRS sampling strategy processes inter-slice information at both sparse and dense intervals to capture comprehensive volumetric context and fine-grained details; (2) CSDHA repurposes existing pre-trained intra-slice attention modules to establish inter-slice information fusion *without adding new parameters*; and (3) SAT extend text-visual alignment to the volumetric domain. Based on BiomedParse [32], extensive evaluation across 11 diverse CT and MRI datasets demonstrates significant improvements in 3D medical image segmentation performance. Furthermore, this design can keep the versatility of working with 2D medical image modalities, and potentially can be applicable to extend other 2D biomedical FMs for facilitating the 3D medical imaging analysis.

2 Method

As shown in Fig. 2, we propose a novel framework to adapt the pre-trained text-prompted 2D biomedical FM BiomedParse [32] for 3D medical image analysis. Each volume and its text prompt first undergo our DRS strategy, extracting robust intra-slice features via a frozen pretrained 2D backbone before fusing inter-slice information through our CSDHA mechanism. The resulting volumetric features are aligned with the text prompt using our SAT loss.

2.1 Dual-rate Sampling Strategy

Accurate medical image segmentation requires understanding both the global anatomical context for organ relationships and boundaries, as well as local fine-grained details for precise delineation [18, 32, 29]. Similar to video understanding where both long-range and short-range temporal dependencies matter [9], medical volumes benefit from multi-scale analysis across slices [24]. Motivated by this observation, we first crop a sub-volume to focus on the region of interest. Given this 3D medical sub-volume $V \in \mathbb{R}^{D\times H\times W}$, where D, H, and W denote depth, height, and width respectively, we apply our dual-rate sampling strategy across the slice dimension (i.e., D dimension) to balance comprehensive volumetric understanding with computational efficiency. Our design



(a) Dual-rate sampling strategy (DRS)

(b) Cross-slice Dual-head Attention (CSDHA)

Fig. 2: Framework of our proposed Bio2Vol. (a) DRS processes volumes at different sampling rates—a sparse rate for capturing global anatomical context and a dense rate for preserving local details; (b) CSDHA repurposes existing pre-trained attention modules to propagate information across slices without introducing additional parameters; and (c) SAT maintains coherence between text prompts and volumetric features.

leverages the pretrained FM for strong spatial representation within intra slice while complementing it with inter-slice understanding through two different sampling rates. The low-rate pathway samples slices sparsely with rate r_s, generating sequence $S_{\text{sparse}} = \{V_{i \cdot r_s}\}_{i=0}^{D/r_s - 1}$ to capture global anatomical context. The dense-rate pathway samples more densely with rate $r_d < r_s$, producing sequence $S_{\text{dense}} = \{V_{j \cdot r_d}\}_{j=0}^{D/r_d - 1}$ to preserve fine-grained details and local transitions. This dual-rate design effectively extends the pretrained 2D model's capabilities to 3D medical image analysis by combining robust intra-slice features with inter-slice contextual information. Ablation study by increasing the sequence length D of the sub-volume enhances the effectiveness of this strategy, confirming the ability of our dual-rate approach to effectively leverage extended volumetric context.

2.2 Cross-slice Dual-head Attention

To effectively integrate information across different slices while maintaining the efficiency of BiomedParse's pre-trained architecture [32], inspired by [16], we propose a parameter-efficient adaptation CSDHA that repurposes its existing attention mechanisms. Unlike previous adaptation methods that require additional adapters [19, 31, 22, 3, 28, 37, 25, 10, 27], our approach reuses the existing attention module without introducing new parameters. Given an input sequence from either pathway $S = \{x_1, x_2, \cdots, x_T\}$ where $x_i \in \mathbb{R}^{H \times W}$, we first extract features using the frozen pre-trained 2D encoder E in the BiomedParse [32]:

$$F = E(S) \in \mathbb{R}^{T \times C}, \tag{1}$$

where T is the sequence length ($T = D/r_s$ for sparse pathway and $T = D/r_d$ for dense pathway) and C is the feature dimension. The feature representation F is then linearly projected using learnable weight matrices W^Q_{attn}, W^K_{attn}, and W^V_{attn} to obtain query, key, and value representations:

$$Q, K, V = FW^Q_{\text{attn}}, FW^K_{\text{attn}}, FW^V_{\text{attn}}. \tag{2}$$

In CSDHA design, we divide the h attention heads into two groups: $h-k$ heads for intra-slice modeling and k heads for inter-slice modeling. For intra-slice heads, we maintain the original attention mechanism for each position tth in the sequence:

$$\text{I-head}_i = \text{Attention}(Q^t_i, K^t_i, V^t_i). \tag{3}$$

For inter-slice heads, we extend the attention to capture relationships between slice at position t and slices at position $t + \Delta t$:

$$\text{C-head}_i = \text{Attention}(Q^t_i, K^{t+\Delta t}_i, V^{t+\Delta t}_i), (\Delta t_i \neq 0). \tag{4}$$

The final output of CSDHA combines both intra-slice and inter-slice information:

$$\text{CSDHA}(F) = \text{Concat}(\text{C-head}_1, \cdots, \text{C-head}_k, \text{I-head}_{k+1}, \cdots, \text{I-head}_h)W^O_{\text{attn}}. \tag{5}$$

CSDHA repurposes existing attention mechanisms to handle both intra-slice features and inter-slice relationships without extra parameters. By controlling slice offsets Δt via DRS sampling, we capture multi-scale dependencies within the volume while preserving 2D feature extraction capabilities. CSDHA replaces all standard attention modules in the architecture.

2.3 Semantic Alignment for Text and Dual-path Visuals

While BiomedParse [32] demonstrates strong capabilities in text-prompted segmentation of 2D medical images—establishing correspondence between textual descriptions and anatomical structures or pathologies—it does not inherently address the alignment between volumetric data and textual descriptions. This represents a critical challenge in medical volume understanding, where the contextual information spans across multiple slices in 3D space. To bridge this gap, we propose a dual-path alignment mechanism that ensures semantic consistency between volumetric and textual representations. Our approach extends BiomedParse [32] by introducing an alignment loss that maximize the similarity between text features and averaging pooled visual features from both pathways:

$$\mathcal{L}_{align} = -(\text{Sim}(\text{Pool}(F^{final}_{sparse}), E_{text}) + \text{Sim}(\text{Pool}(F^{final}_{dense}), E_{text})), \tag{6}$$

where F^{final}_{sparse}, F^{final}_{dense} represent the final features from sparse and high pathways respectively (*i.e.,* global contexts and local details), E_{text} denotes the text features, Pool averages spatial dimensions, and Sim denotes cosine similarity.

Overall Objective. The total loss combines this novel alignment objective with the original BiomedParse losses [32], *i.e.,* $\mathcal{L} = \mathcal{L}_{\text{biomedparse}} + \lambda\mathcal{L}_{\text{align}}$ where λ is a hyperparameter that balances the two loss components

3 Experiment Results

Implementation Details. We implemented our methodology using Python 3.9, PyTorch 2.4, and MONAI 1.4, running all experiments on an NVIDIA RTX 3090Ti GPU. Following [32], we preprocessed each volumetric dataset by cropping sub-volumes of depth d and resizing the extracted slices to 1024×1024. The preprocessing pipeline included intensity thresholding and max-min normalization to the [0, 1] range. We fintuned the model [32] using Adam optimizer [6] with a learning rate of 1e-5 and batch size of 5 for 20 epochs. The alignment loss weight λ was set to 1, with $r_d = 1$ and $r_s = 2$ and $d = 7$. For CSDHA, we configured $h = 12$, and empirically set $k = 1$ and $\Delta t = 2$. We evaluated model performance using three metrics: Dice score coefficient (DSC), Normalized Surface Dice (NSD), and 95th percentile Hausdorff Distance (HD95).

Datasets. We evaluated Bio2Vol on 11 public 3D medical image datasets across CT and MRI modalities, following the partitioning in [32]. These include the Amos22 dataset [14] (500 CT scans for abdominal organ segmentation) and the Medical Segmentation Decathlon (MSD) [1], which contains ten distinct organ and tumor segmentation tasks. More details are available in [32]. We followed [32] to construct the text prompt for each volume data.

Table 1: Comparison of Dice scores (%) on the Amos22 dataset [14], with best and second-best results bolded and underlined, respectively. Per-class standard deviations are omitted for brevity. Methods marked with † adapt 2D biomedical foundation models to 3D medical imaging. ($p < 0.05$)

Methods	Dice score(%)															**AVG**
	LI	ST	SP	LK	PA	IV	AO	RK	BL	DU	ES	GB	LA	RA	PR	
3D backbone																
nnUnet[13]	90.86	91.21	89.47	78.42	73.18	94.98	62.84	89.12	61.90	75.55	60.96	57.97	65.98	74.47	66.58	$78.88_{\pm 0.97}$
UNETR[12]	89.51	90.61	85.49	72.81	76.05	93.44	78.62	90.05	81.59	76.99	65.27	62.19	62.37	69.76	69.61	$77.62_{\pm 0.04}$
nnFormer[35]	93.31	92.55	92.29	83.10	75.11	95.91	88.21	89.39	83.69	80.19	63.41	59.99	70.92	78.48	73.96	$81.37_{\pm 0.69}$
SegMamba[30]	90.66	93.43	91.47	82.86	79.34	94.85	86.66	91.87	87.01	81.32	69.26	64.95	73.40	74.73	75.61	$82.49_{\pm 2.42}$
SwinUNETR[11]	91.28	94.20	92.86	82.40	80.02	95.72	88.36	91.27	87.25	81.73	68.55	68.91	72.75	**79.99**	**76.05**	$83.42_{\pm 0.34}$
Visual-prompt based																
SAM [15]	68.20	61.47	74.07	80.43	39.93	52.77	77.33	79.93	61.83	38.13	49.10	69.97	31.73	24.33	68.32	$58.50_{\pm 10.12}$
3DSAM-adapter† [10]	86.60	81.23	91.20	88.57	71.87	85.80	92.20	87.90	89.27	80.03	81.33	87.37	74.43	60.70	69.03	$81.84_{\pm 1.35}$
MA-SAM† [3]	94.33	86.80	88.30	**91.67**	79.10	82.00	89.70	91.67	89.79	80.27	79.40	74.37	74.36	68.73	69.12	$82.64_{\pm 1.06}$
Medical SAM2† [37]	91.23	85.70	90.17	83.10	80.70	90.17	86.83	88.90	83.13	83.37	83.87	77.48	79.12	67.36	67.35	$82.57_{\pm 0.53}$
Text-prompt based																
BiomedParse[32]	92.59	93.17	90.23	82.10	74.82	88.60	86.35	90.50	88.45	85.21	65.22	61.55	77.10	66.69	68.35	$80.73_{\pm 0.70}$
Ensemble† [34]	94.55	93.67	91.46	82.46	76.31	88.51	86.82	90.41	88.19	85.02	66.45	62.78	77.44	68.26	69.45	$81.45_{\pm 0.24}$
ST-Apater† [19]	95.73	95.10	95.23	65.27	82.27	**97.03**	90.67	90.30	87.30	78.87	**86.43**	**74.33**	77.45	68.07	69.49	$83.40_{\pm 0.34}$
AIM† [31]	95.13	94.93	94.50	85.30	82.27	95.57	89.90	92.70	87.27	83.37	77.87	72.77	74.67	69.67	68.97	$84.32_{\pm 0.59}$
Ours	**96.20**	**96.50**	**95.90**	84.40	**85.60**	96.40	**94.30**	**94.10**	**92.40**	**85.40**	70.20	72.50	**79.13**	70.50	75.00	$\mathbf{85.45}_{\pm 0.45}$

Note: LI (Liver), ST (Stomach), SP (Spleen), LK (Left Kidney), PA (Pancreas), IV (Inferior Vena Cava), AO (Aorta), RK (Right Kidney), BL (Bladder), DU (Duodenum), ES (Esophagus), GB (Gall Bladder), LA (Left Adrenal Gland), RA (Right Adrenal Gland), PR (Prostate).

Performance Comparison. As shown in Table 1, 3D backbone segmentation methods demonstrate competitive performance by processing volumetric data holistically. Direct application of 2D visual-based models like SAM [15] yields suboptimal results due to missing volume context. Adaptation approaches such as 3DSAM-adapter [10] and MA-SAM [3] significantly improve performance through spatial and temporal adapters. Medical SAM2 [25], despite being a 2D FM, achieves comparable results by using memory blocks to model inter-slice relationships with proper fine-tuning, indicating the importance of learning volume context. Recent text-prompt-based 2D biomedical FMs BiomedParse [32] also show strong performance from training on large, diverse medical datasets across nine modalities. Parameter-efficient tuning methods (Rows 11-13) in BiomedParse [32] demonstrate substantial improvements by effectively modeling continuous slice information. Notably, text-prompted adaptation methods consistently outperform visual-prompt SAM-based approaches. Our proposed method achieves mean Dice score of 85.45% and statistically significant improvement ($p < 0.05$) over existing approaches. This improvement stems from better modeling of inter-slice relationships while preserving learnable features and strengthening the correlation between volume context and text prompts. Fig. 3 provides qualitative comparisons of segmentation outcomes.

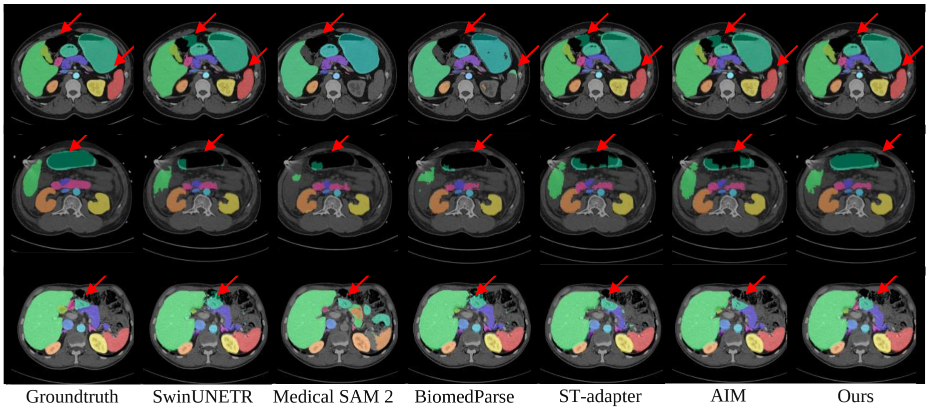


Fig. 3: Qualitative comparison of abdominal organ segmentation results on the Amos22 dataset [14] using the text prompt: “Organs in the abdomen in the CT.”

Generalization Analysis. We evaluated Bio2Vol’s generalization capability on the MSD dataset [1]. As shown in Table 2, Bio2Vol consistently outperforms BiomedParse [32] and SOTA methods across all metrics (DSC, NSD, and HD95). Notable improvements appear in challenging tasks like Task03 and Task05, particularly in HD95 measurements. By leveraging continuous slice information, Bio2Vol better models 3D anatomical structures compared to 2D FMs that process slices independently. While AIM [31] introduced joint spatial-temporal adapters for inter-slice modeling, our approach achieves significantly better results through two complementary innovations: 1) the CSDHA module paired with our DRS strategy, which together enhance volumetric context modeling; and 2) our SAT module that strengthens the alignment between volumetric features and text prompts. This parameter-efficient design enables more comprehensive modeling of anatomical structures across multiple slices while maintaining the FM’s original strengths, allowing better represent 3D anatomical structures.

Ablation Study. As shown in Table 3, the base model [32] (*i.e.,* BiomedParse without adaptation), achieves a DSC of 80.73% and NSD of 78.25% on the Amos22 dataset [14]. Adding CSDHA improves the performance to 85.01% DSC and 79.32% NSD, demonstrating its effectiveness in capturing inter-slice

Table 2: Quantitative comparison with SOTA adaptation methods based on BiomedParse [32] across ten tasks on MSD dataset [1].

Metrics	Method	Task01	Task02	Task03	Task04	Task05	Task06	Task07	Task08	Task09	Task10
DSC↑	BiomedParse [32]	80.54$_{\pm 0.63}$	86.51$_{\pm 1.85}$	78.68$_{\pm 1.44}$	81.24$_{\pm 1.67}$	25.75$_{\pm 0.89}$	11.06$_{\pm 0.64}$	46.23$_{\pm 1.12}$	43.84$_{\pm 1.34}$	60.71$_{\pm 1.56}$	61.33$_{\pm 0.72}$
	AIM† [31]	81.33$_{\pm 1.45}$	86.95$_{\pm 1.88}$	79.42$_{\pm 1.52}$	82.45$_{\pm 1.75}$	26.89$_{\pm 0.91}$	12.85$_{\pm 0.67}$	47.38$_{\pm 1.18}$	44.65$_{\pm 1.37}$	61.54$_{\pm 1.58}$	61.98$_{\pm 1.49}$
	Ours	**83.52**$_{\pm 0.59}$	**87.98**$_{\pm 1.92}$	**81.32**$_{\pm 1.65}$	**84.22**$_{\pm 1.88}$	**29.32**$_{\pm 0.95}$	**16.32**$_{\pm 0.72}$	**49.12**$_{\pm 1.28}$	**46.32**$_{\pm 1.41}$	**63.12**$_{\pm 1.62}$	**63.12**$_{\pm 1.58}$
NSD↑	BiomedParse [32]	69.60$_{\pm 0.75}$	81.69$_{\pm 1.76}$	36.48$_{\pm 0.98}$	37.13$_{\pm 1.05}$	24.80$_{\pm 0.85}$	5.55$_{\pm 0.52}$	42.43$_{\pm 1.15}$	64.19$_{\pm 1.48}$	64.19$_{\pm 1.52}$	59.53$_{\pm 1.38}$
	AIM† [31]	70.86$_{\pm 0.82}$	82.15$_{\pm 1.78}$	37.42$_{\pm 1.01}$	39.25$_{\pm 1.12}$	25.95$_{\pm 0.87}$	7.89$_{\pm 0.58}$	43.48$_{\pm 1.17}$	64.78$_{\pm 1.51}$	64.48$_{\pm 1.53}$	60.42$_{\pm 1.41}$
	Ours	**73.40**$_{\pm 0.56}$	**83.12**$_{\pm 1.82}$	**39.23**$_{\pm 1.08}$	**43.22**$_{\pm 1.25}$	**28.12**$_{\pm 0.92}$	**12.65**$_{\pm 0.68}$	**45.32**$_{\pm 1.22}$	**66.12**$_{\pm 1.58}$	**65.12**$_{\pm 1.55}$	**62.12**$_{\pm 1.46}$
HD95↓	BiomedParse [32]	32.78$_{\pm 1.96}$	18.21$_{\pm 0.75}$	303.44$_{\pm 1.95}$	46.19$_{\pm 1.32}$	85.89$_{\pm 1.72}$	291.51$_{\pm 1.94}$	75.43$_{\pm 1.64}$	65.35$_{\pm 1.51}$	213.17$_{\pm 1.89}$	77.44$_{\pm 1.68}$
	AIM† [31]	30.94$_{\pm 0.93}$	17.35$_{\pm 0.73}$	290.16$_{\pm 1.94}$	45.82$_{\pm 1.30}$	83.85$_{\pm 1.71}$	278.42$_{\pm 1.93}$	73.65$_{\pm 1.62}$	66.12$_{\pm 1.52}$	204.86$_{\pm 1.88}$	76.67$_{\pm 1.67}$
	Ours	**26.12**$_{\pm 2.55}$	**15.32**$_{\pm 0.71}$	**264.12**$_{\pm 1.93}$	**45.01**$_{\pm 1.28}$	**80.12**$_{\pm 1.69}$	**250.32**$_{\pm 1.91}$	**70.12**$_{\pm 1.59}$	**67.12**$_{\pm 1.53}$	**190.12**$_{\pm 1.86}$	**75.12**$_{\pm 1.65}$

continuous relationships. The integration of DRS further enhances the model's capabilities, reaching 85.38% DSC and 81.85% NSD by leveraging both local and global contextual information. Finally, incorporating SATV provides better alignment between text and volume features, achieving our best performance of 85.45% DSC and 82.17% NSD. These improvements are consistent across the MSD Task01 brain tumor MRI dataset [1], where we observe similar progressive enhancements from baseline (80.54% DSC, 69.60% NSD) to our full model (83.52% DSC, 73.40% NSD). Our analysis of sub-volume depth d (Fig.4) shows that performance improves rapidly as we increase the number of slices used for modeling inter-slice relationships. While increasing sub-volume depth improves performance, it adds computational cost. However, our method avoids introducing extra parameters like adapters [31, 19, 3, 25], resulting in lower overhead. As shown in Table. 4, the optimal numbers of inter-slice head k would be 1.

Table 3: Ablation study of key modules on Amos22 dataset [14] and MSD Task01 dataset [1].

Dataset	Base	CSDHA	DRS	SATV	DSC(%)↑	NSD(%)↑	HD95↓
Amos22 [14]	✓				80.73±0.35	78.25±0.45	88.57±3.35
	✓	✓			85.01±0.38	80.32±0.33	68.12±2.10
	✓	✓	✓		85.38±0.43	81.85±0.48	54.85±3.40
	✓	✓	✓	✓	**85.45**±0.45	**82.17**±0.50	**52.17**±3.55
MSD Task01 [1]	✓				80.54±0.64	69.60±0.65	32.78±2.85
	✓	✓			82.84±0.61	71.20±0.60	30.18±1.55
	✓	✓	✓		83.05±0.49	72.85±0.58	27.45±2.40
	✓	✓	✓	✓	**83.52**±0.59	**73.40**±0.56	**26.12**±2.55

Table 4: Analysis of inter-slice head k on Amos22.

k	0	1	2	4
DSC(%)	82.15±0.65	**85.45**±0.45	85.30±0.43	84.31±0.41

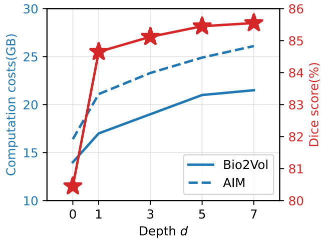


Fig. 4: Performance and computation costs on the Amos22 dataset [14].

4 Conclusion

In this paper, we present **Bio2Vol**, a novel framework for adapting BiomedParse [32] FMs to text-prompted 3D medical image segmentation. Our approach bridges the gap between 2D biomedical foundation models and volumetric data through three key innovations. Comprehensive evaluation across 11 CT and MRI datasets demonstrates that Bio2Vol improves segmentation accuracy. Furthermore,our approach maintains computational efficiency without introducing extra parameter for adoptatopm while preserving BiomedParse's sophisticated 2D pretrained capabilities. This work establishes an effective paradigm for adapting 2D biomedical foundation models to 3D medical image segmentation, with potential applications to other 2D foundation models and medical imaging tasks.

Acknowledgments. This work is supported

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

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nat. Commun **13**(1), 4128 (2022)
2. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv (2024)
3. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. MIA **98**, 103310 (2024)
4. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. MIA **97**, 103280 (2024)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
6. Diederik, P.K.: Adam: A method for stochastic optimization. ICLR (2014)
7. Du, Y., Zhuang, J., Zheng, X., Cong, J., Guo, L., He, C., Luo, L., Li, X.: Beyond h&e: Unlocking pathological insights with polarization via self-supervised learning. arXiv (2025)
8. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. NeurIPS **37**, 110746–110783 (2025)
9. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV. pp. 6202–6211 (2019)
10. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. MIA **98**, 103324 (2024)
11. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: MICCAIW. pp. 272–284 (2021)
12. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR : Transformers for 3d medical image segmentation. In: WACV. pp. 574–584 (2022)
13. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
14. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. NeurIPS **35**, 36722–36732 (2022)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: CVPR. pp. 4015–4026 (2023)
16. Li, X., Zhu, Y., Wang, L.: Zeroi2v: Zero-cost adaptation of pre-trained transformers from image to video. In: ECCV. pp. 425–443 (2024)

17. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: ICCV. pp. 21152–21164 (2023)
18. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun **15**(1), 654 (2024)
19. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. NeurIPS **35**, 26462–26477 (2022)
20. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: MICCAI. pp. 405–415. Springer (2023)
21. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. **19**(1), 221–248 (2017)
22. Shen, Y., Li, J., Shao, X., Inigo Romillo, B., Jindal, A., Dreizin, D., Unberath, M.: Fastsam3d: An efficient segment anything model for 3d volumetric medical images. In: MICCAI. pp. 542–552 (2024)
23. Sun, Y., Wang, L., Li, G., Lin, W., Wang, L.: A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. Nat. Biomed. Eng. pp. 1–18 (2024)
24. Wu, B., Xiao, Q., Liu, S., Yin, L., Pechenizkiy, M., Mocanu, D.C., Keulen, M., Mocanu, E.: E2enet: Dynamic sparse feature fusion for accurate and efficient 3d medical image segmentation. NeurIPS **37**, 118483–118512 (2025)
25. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv (2023)
26. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: CVPR. pp. 22873–22882 (2024)
27. Xiang, W., Li, C., Wang, B., Wei, X., Hua, X.S., Zhang, L.: Spatiotemporal self-attention modeling with temporal patch shift for action recognition. In: ECCV. pp. 627–644. Springer (2022)
28. Xie, B., Tang, H., Duan, B., Cai, D., Yan, Y.: Masksam: Towards auto-prompt sam with mask classification for medical image segmentation. arXiv (2024)
29. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: MICCAI. pp. 171–180 (2021)
30. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: MICCAI. pp. 578–588 (2024)
31. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: Aim: Adapting image models for efficient video action recognition. ICLR (2023)
32. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moung-Wen, C., Piening, B., Bifulco, C., Wei, M., Poon, H., Wang, S.: A foundation model for joint segmentation, detection, and recognition of biomedical objects across nine modalities. Nature Methods (2024)
33. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: One model to rule them all: Towards universal segmentation for medical images with text prompts. arXiv (2023)
34. Zheng, J., Cao, X., Zhang, B., Zhen, X., Su, X.: Deep ensemble machine for video classification. TNNLS **30**(2), 553–565 (2018)
35. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv (2021)
36. Zhou, T., Li, L., Bredell, G., Li, J., Unkelbach, J., Konukoglu, E.: Volumetric memory network for interactive medical image segmentation. MIA **83**, 102599 (2023)

37. Zhu, J., Qi, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv (2024)
38. Zhuang, J., Wu, L., Wang, Q., Fei, P., Vardhanabhuti, V., Luo, L., Chen, H.: Mim: Mask in mask self-supervised pre-training for 3d medical image analysis. TMI (2025)