TAGS: 3D Tumor-Adaptive Guidance for SAM

Sirui Li[1] Linkai Peng[2] Zheyuan Zhang[2] Gorkem Durak[2] Ulas Bagci[2]
[1]Southern University of Science and Technology [2]Northwestern University
12012719@mail.sustech.edu.cn,
{linkai.peng, zheyuan.zhang, gorkem.durak, ulas.bagci}@northwestern.edu

Abstract

*Foundation models (FMs) such as CLIP and SAM have recently shown great promise in image segmentation tasks, yet their adaptation to 3D medical imaging—particularly for pathology detection/segmentation—remains underexplored. A critical challenge arises from the domain gap between natural images and medical volumes: existing FMs, pre-trained on 2D data, struggle to capture 3D anatomical context, limiting their utility in clinical applications like tumor segmentation. To address this, we propose an adaptation framework called **TAGS**: **T**umor **A**daptive **G**uidance for **S**AM, which unlocks 2D FMs for 3D medical tasks through multi-prompt fusion. By preserving most of the pre-trained weights, our approach enhances SAM's spatial feature extraction using CLIP's semantic insights and anatomy-specific prompts. Extensive experiments on three open-source tumor segmentation datasets prove that our model surpasses the state-of-the-art medical image segmentation models (+46.88% over nnUNet), interactive segmentation frameworks, and other established medical FMs, including SAM-Med2D, SAM-Med3D, SegVol, Universal, 3D-Adapter, and SAM-B (at least +13% over them). This highlights the robustness and adaptability of our proposed framework across diverse medical segmentation tasks. Our code and model are available at: https://github.com/sirileeee/TAGS.*

1. Introduction

Medical image segmentation is essential for applying advanced Artificial Intelligence (AI) technologies in clinical settings, increasing diagnostic accuracy, and optimizing personalized treatment planning [49, 51]. Tumor segmentation, in particular, remains one of the most challenging tasks due to the ambiguous boundaries and varying sizes and locations of tumors, along with other unique challenges related to scanners and patient population differences [23, 37].

While convolutional neural networks (CNNs), particularly U-Net [42] and its variants [36, 59], have dominated

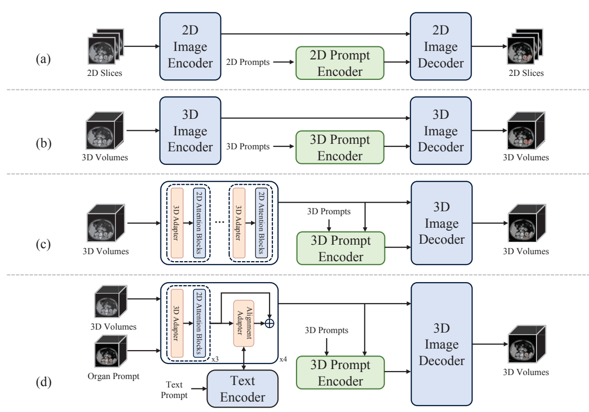

Figure 1. Comparison of 4 approaches for SAM-driven volumetric tumor segmentation: (a) slice-by-slice segmentation using 2D models; (b) employing 3D medical foundation models; (c) adapting 2D models for 3D segmentation; (d) our proposed TAGS, featuring tumor-specific adaptations.

medical segmentation by leveraging localized feature extraction, their limited capacity to model global contextual relationships often leads to suboptimal delineation of tumors from surrounding tissues. In contrast, Transformer-based architectures [48] address this issue through self-attention mechanisms to model long-range dependencies, effectively capturing global context correlations. Thus, they have demonstrated their strengths in tumor segmentation [13, 19, 38], while still with weaknesses: their computational complexity and reliance on large-scale annotated datasets hinder practical deployment in clinical workflows.

Recently, transformer-based foundation models [30, 34, 57] have made substantial breakthroughs in computer vision, particularly in segmentation tasks through promptable inference and generalization to unseen data. With large-scale pre-training, these models perform better across diverse downstream tasks and handle unseen data more efficiently. The Segment Anything Model (SAM) [24] exemplifies this trend with its interactive object-prompting mechanism. However, studies have shown that applying SAM directly to medical image segmentation presents unique challenges [4, 35]. First, medical images are typically 3D volumetric data, while SAM is pre-trained on 2D images, limiting its capacity to capture 3D spatial information effectively. Although SAM2 [41], designed for video segmentation, provides additional contextual dimensions, it remains insufficient for medical segmentation tasks [44]. Furthermore, the significant domain gap between natural and medical images constrains SAM's ability to generalize effectively to medical data.

However, applying SAM directly to 3D medical imaging—particularly tumor segmentation—faces two critical limitations [4, 35]. First, SAM's 2D pre-training on natural images fails to capture the inter-slice spatial relationships inherent to volumetric medical data, resulting in inconsistent predictions across adjacent slices. Although SAM2 [41], designed for video segmentation, provides an additional contextual (temporal) dimension, it remains insufficient for medical segmentation tasks [44]. Second, the semantic-contextual misalignment between natural and medical domains—such as the hierarchical dependency between organs and tumors—undermines SAM's ability to localize pathology accurately.

Prior methods have explored several approaches, as illustrated in Fig. 1. One approach is slice-by-slice segmentation, which decouples volumetric images into 2D slices [9, 27, 33]. It applies SAM to each slice individually and then stacks the segmentation results to form a 3D output. This method, however, requires extensive prompting for each slice and fails to capture inter-slice correlations, which leads to inconsistent predictions across the volume. Another approach aims to develop 3D medical segmentation foundation models [15, 50]. Nonetheless, the scarcity of publicly accessible 3D medical datasets with tumor annotations limits these models. As a result, they are often pre-trained for both organ and tumor segmentation together. Yet, the difficulty in aligning data from diverse sources and the high computational demands for pre-training further limit this approach's feasibility.

Incorporating 3D adapters into the original SAM architecture has also proven effective in this manner [6, 16, 54]. By adding 3D adapters, the FM captures spatial information while retaining the majority of SAM's pre-trained weights. Throughout the process, SAM's 2D encoders stay frozen, which minimizes the number of trainable parameters. This setup accelerates both model development and deployment, enhancing application speed. However, the lack of spatial awareness in the original pre-trained SAM can limit its ability to fully capture volumetric features.

To address all these challenges, we propose TAGS, a 3D **T**umor-**A**daptive **G**uidance for **S**AM, a parameter-efficient framework that adapts SAM's 2D capabilities to 3D tumor segmentation through synergistic integration of CLIP's semantic priors [40] and anatomy-specific prompting. Unlike prior works that naively fuse CLIP's feature maps with SAM [25, 29], TAGS employs a hierarchical prompting mechanism where CLIP's text-guided embeddings refine SAM's spatial attention to prioritize anatomically relevant regions. Pixel-wise loss tackles the foreground-background pixel imbalance in volumetric tumor segmentation, optimizing learning efficiency. Our organ-specific prompts, generated via automated segmentation tools, constrain SAM's focus to tumor-prone anatomical contexts, mitigating false positives caused by spatial uncertainty.

Our major contributions are threefold:

- We propose the first multi-level feature alignment framework to unify SAM's spatial acuity with CLIP's semantic embeddings for 3D tumor segmentation, requiring only 18% of SAM-B's tunable parameters. This hierarchical prompting mechanism bridges the 2D-to-3D domain gap while preserving computational efficiency.
- To overcome the spatial uncertainty of tumors, we design an organ-specific prompt module that dynamically localizes tumors within anatomically plausible regions, directing the model to segment tumors within specific organ regions and reduce false positives.
- Extensive experiments on multiple datasets (Liver, Kidney, and Pancreas Tumors) consistently validate our approach, achieving new state-of-the-art results in volumetric tumor segmentation, with an average Dice score improvement of +46.88% over nnUNet [21] and at least +13% over medical FMs [15, 50].

2. Related Work

2.1. Volumetric Tumor Segmentation

Volumetric tumor segmentation is fundamental for oncology workflows, including accurate cancer diagnosis, staging, and prognosis. In comparison to 2D segmentation, volumetric (3D) approaches must overcome increased computational overhead, higher data dimensionality, and the complexity of delineating ambiguous tumor boundaries across multiple slices. Classic 3D methods such as 3D U-Net [11], TransUnet [7], UNetr [17], and Cotr [55] have set benchmarks for volumetric organ segmentation. However, their performance on tumors remains suboptimal, largely due to irregular tumor shapes, low contrast in medical scans, and diffuse boundaries. Even foundation models (FMs) trained on 3D medical images from scratch [15, 50] show limited performance due to insufficient volumetric tumor annotations for pre-training and the high variability in tumor size, texture, and shape compared to organs. Addressing these challenges, this work advances tumor-specific 2D-to-3D adaptation to achieve more precise volumetric delineation.

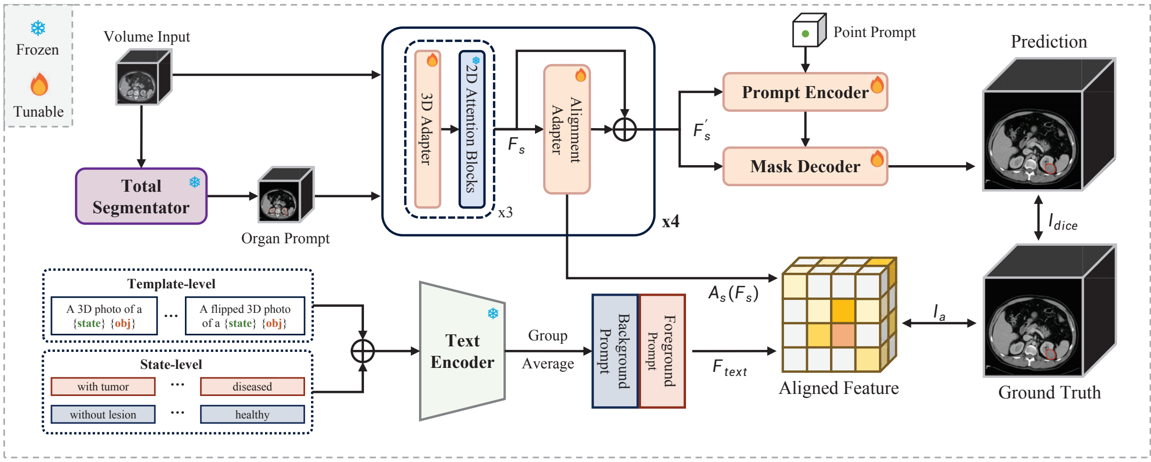

Figure 2. The overall framework of proposed model (TAGS). TAGS includes 3 different prompts: (a) annotation-free organ prompt; (b) dual-category text prompt generated by CLIP text encoder; (c) interactive 3D point prompts. Organ prompt is input into the image encoder with the raw image. Stage-wise alignment adapters then integrate foreground and background semantic information from the text prompt into the image feature extraction. Finally, the image features merge with the point prompt, and the mask decoder predicts the tumor region.

2.2. 2D-to-3D Foundation Model Adaptation

Training new 3D foundation models from scratch requires a vast amount of 3D data, which is often scarce and computationally more expensive to process. This prohibitive cost of training 3D foundation models from scratch has spurred interest in adapting 2D pre-trained models for volumetric tasks. Specifically, the abundance of high-quality 2D data [43, 47, 58] has driven rapid innovation in 2D foundation models, prompting efforts to adapt them for volumetric tasks. One prominent strategy involves injecting 3D adapters into a frozen 2D image encoder [6, 16]. By inflating 2D convolutions, these methods capture volumetric information with minimal trainable parameters. Analogous to image-to-video adaptation, [54] uses space-depth adapters to incorporate 3D context. These approaches confirm that strong 3D performance is achievable by reusing the extensive knowledge embedded in 2D backbone weights. In medical imaging, such adaptation is particularly advantageous, where data heterogeneity, scanner variations, and high annotation costs complicate end-to-end 3D training.

2.3. Visual-Language Alignment

Studies have shown that visual-language models (VLM), FMs in other words, which embed semantic information into visual learning, surpass vision-only models in tasks like classification [12], segmentation [61], and anomaly detection [28]. CLIP [40], pre-trained on vast image-text datasets, achieves a good generalization ability in classification tasks, even effectively identifying previously unseen samples. Specifically in medical AI, MedCLIP [52], trained on extensive medical image-text data, excels in zero-shot prediction, supervised classification, and image-text retrieval. While [39] and [1] fine-tuned CLIP for 2D segmentation via cross-modal attention, recent work [32] extended this to 3D by fusing CLIP's embeddings with 3D CNN features, but retained a modality-specific design where text guidance is applied only during decoding. In contrast, we propose **hierarchical vision-language prompting** that injects CLIP's semantic priors directly into SAM's encoder, enabling earlier fusion of anatomical context. This approach differs fundamentally from prior VLMs, which treat text as a post-hoc classifier, and aligns with clinical workflows where radiological reports guide attention during image interpretation.

3. Methodology

3.1. Preliminaries

We build upon two key components: SAM [24] and 3D SAM Adapter [16]. SAM is a versatile, promptable segmentation model that that leverages a ViT-based [14] encoder-decoder architecture and a prompt encoder, allowing interactive segmentation via points, boxes, or masks. Although pre-trained on extensive 2D image datasets, SAM faces two primary challenges in medical imaging: (1) the lack of domain-specific data in its pre-training leads to domain shift and reduced robustness, and (2) its 2D-centric design struggles to capture volumetric structure, typical of medical scans.

To efficiently apply SAM to 3D tumor segmentation, Gong *et al.* [16] introduced the **3D SAM Adapter**, retaining SAM's pre-trained encoder weights while converting its 2D operations into 3D. Specifically, 2D convolutions are replaced by 3D convolutions, and an additional dimension is added to patch embeddings, positional encodings, and attention queries, forming a 3D ViT. A lightweight spatial adapter, comprising projection and 3D convolution layers, is inserted between consecutive attention blocks to facilitate 3D feature learning. Only the convolution, spatial adapter, and normalization parameters are fine-tuned, while all other parameters remain frozen—ensuring memory efficiency and mitigating catastrophic forgetting. A multi-layer aggregation decoder [60] concatenates the encoder's intermediate outputs to produce segmentation masks. Despite these adaptations, the 3D SAM Adapter [16] supports only point-based prompts, which contain limited spatial context [62] to align prompt and image embeddings. Therefore, accuracy can still suffer in volumetric segmentation tasks.

To overcome these limitations, we propose **TAGS**, a 3D tumor-adaptive guidance framework that leverages multiple prompt types for improved tumor segmentation. Fig. 2 illustrates the overall architecture for our proposed framework.

3.2. Annotation-free Organ Prompt

Tumor segmentation is complicated by variability in size, location, and appearance. However, tumors generally reside within specific organs. Rather than multitask training for simultaneous organ and tumor segmentation [10, 56], we propose a simpler method that circumvents the need for organ ground-truth labels.

As depicted in the top-left of Fig. 2, we use an automated tool, *TotalSegmentator* [53], to generate pseudo-organ masks $M \in \mathbb{R}^{d\times h\times w}$, where d, h, and w are the volume (input) dimensions. These masks are typically more accurate and less costly to obtain than tumor annotations. To align with SAM's three-channel expectation, we replicate the original volume and replace its third channel with the organ mask, forming $I \in \mathbb{R}^{3\times d\times h\times w}$. This overlay provides organ-level spatial context, guiding the model's attention to relevant anatomical regions and improving tumor feature extraction.

3.3. Multi-level Feature Alignment

Dual-category Text Prompt. Since tumors often occupy $<$1% of the entire organ volume, relying solely on a single "tumor" description is insufficient. Inspired by [22] and [20], we adopt a hierarchical dual-category text prompting strategy that encodes both foreground (tumor) and background (healthy tissue). Each prompt includes high-level "state" descriptors (e.g., "lesion present/absent") and generic "template" descriptions, as shown in Fig. 2. Only the organ's name is required, reducing dependence on specialized medical text encoders. To minimize bias, we generate multiple text prompts for each category (see supplementary material). We then aggregate their features with CLIP's [40] text encoder to yield $F_{text} \in \mathbb{R}^{c\times 2}$, where c is the feature dimension.

Stage-wise Alignment Adapter. SAM's image encoder comprises 12 attention blocks (Fig. 2) grouped into four stages (three blocks per stage). We insert a lightweight adapter A_s between stages to align image and text features more granularly. Each alignment adapter is designed with a single fully connected layer to minimize the number of parameters. Each input volume I' passes through four stages with the assistance of a 3D adapter to generate visual features. Intermediate features at each stage are denoted as $F_s \in \mathbb{R}^{P_d\times P_h\times P_w\times c}$, $s \in \{1,2,3,4\}$. Here, P_d, P_h, and P_w represent the number of patches in the depth, height, and width dimensions after patchifying the original image. The alignment adapter A_s could be formulated as:

$$A_s(F_s) = \sigma(F_s W_s),\, s \in \{1,2,3,4\}, \tag{1}$$

where $\sigma(\cdot)$ is the activation function and W_s indicates the parameter of A_s. We also adopt residual connection to enhance robustness. With a scaling factor $\lambda = 0.2$, the final output of stage s, denoted as $F_s^{'}$, is defined as:

$$F_s^{'} = \lambda A_s(F_s) + (1-\lambda)F_s,\, s \in \{1,2,3,4\}. \tag{2}$$

Multi-modal Alignment Loss. We incorporate tumor annotations into a multi-modal alignment loss. First, we calculate the *cosine similarity* ($sim(\cdot,\cdot)$) for each adapter output $A_s(F_s)$. After interpolation and softmax, we obtain dense predictions. Given the severe foreground-background imbalance in volumetric tumor segmentation, we employ a focal loss [31] combined with dice loss [5]:

$$\begin{aligned} l_a = \sum_{s=1}^{4}\{\frac{1}{2} l_{focal}[\phi(sim(A_s(F_s), F_{text}), y)] + \\ \frac{1}{2} l_{dice}[\phi(sim(A_s(F_s), F_{text}), y]\}, \end{aligned} \tag{3}$$

where ϕ refers to interpolation and *softmax*, and y is the tumor ground truth. Although these predictions are dense, the final segmentation still requires the decoder with point-based prompts. Experimental results in Sec. 4.3 verify this.

3.4. Generating Final Segmentations

We follow the prompt encoder and decoder structure of [16]. During training, we randomly sample $n = 10$ points from foreground and background if a tumor is present; otherwise, all $n = 10$ points come from the background. The decoder concatenates intermediate encoder outputs $F_s^{'}$ with the model input I' to produce the segmentation $\hat{y}$. We optimize a dice loss on $\hat{y}$:

$$L = l_{dice}(\hat{y}, y) + l_a. \tag{4}$$

At inference, we preserve the organ prompt and alignment adapters but **omit** the text prompt. We randomly pick 1 or 3 points inside the tumor region as input to the prompt encoder. The image is cropped around these points to match training patch sizes, and the encoder-decoder stack outputs the final mask. Experimental results in Sec. 4.3 show that text-based alignment in training, combined with targeted

Category	Method	Kidney Tumor		Liver Tumor		Pancreas Tumor		Tunable
		Dice (%)	NSD (%)	Dice (%)	NSD (%)	Dice (%)	NSD (%)	Parameters
Classic	nnUNet [21]	50.24	47.62	58.95	57.58	36.46	43.18	30.76 M
	3D UX-Net [26]	53.40	57.02	47.95	61.14	32.11	48.46	53.01 M
	Swin-UNETR [46]	49.83	54.06	43.90	61.39	35.61	53.96	62.19 M
	UNETR++ [45]	54.15	55.09	35.33	47.42	21.50	37.22	55.70 M
2D SAM-based	SAM-B [24]	4.35	4.34	2.96	3.49	4.75	5.65	-
	SAM-Med2D [8]	24.57	29.86	12.43	17.70	17.78	27.15	-
3D SAM-based	SAM-Med3D [50]	27.32	36.49	10.37	17.01	16.28	21.90	-
	SAM-Med3D Turbo [50]	47.83	66.40	19.34	30.95	34.64	57.62	-
	SegVol [15]	27.44	30.33	40.92	31.72	54.10	64.56	-
	SegVol w zoom [15]	28.14	31.51	**66.93**	64.62	58.60	66.03	-
CLIP-based	Universal [32]	74.95	73.50	61.87	60.72	**61.28**	70.51	-
2D-to-3D	3D Adapter (1pts) [16]	76.67	81.95	52.80	69.46	55.02	77.92	25.46 M
	TAGS (1pts)	<u>80.39</u>	<u>87.69</u>	59.69	72.83	59.96	<u>82.05</u>	27.82 M
	3D Adapter (3pts) [16]	77.16	82.76	58.96	<u>73.46</u>	55.73	78.24	25.46 M
	TAGS (3pts)	**80.83**	**88.26**	<u>66.23</u>	**79.33**	<u>61.04</u>	**83.10**	27.82 M

Table 1. The comparison experiments between TAGS and other benchmarks. For SAM-based foundation models, we use them directly for inference, while other models are trained on the training set. The best results are **bold** and the second best ones are <u>underlined</u>.

prompts at test time, significantly enhances tumor segmentation accuracy. One to three point prompts balance usability and accuracy in volumetric segmentation. Using only one point offers a quick, minimal input requirement, making the approach more feasible in busy clinical settings. Meanwhile, adding two additional points (for a total of three) provides more geometric coverage within the tumor region, improving localization and boundary delineation. Empirical testing demonstrated that beyond three points, gains in segmentation accuracy diminish while user effort grows, so we chose to optimize performance without overburdening the prompt-creation process.

4. Experiments and Results

4.1. Setup

Datasets and Evaluation Metrics. We focus exclusively on tumor segmentation—arguably the most challenging task for SAM-based methods in medical imaging. Three public datasets, each targeting a different tumor type, are used: (a) the kidney tumor segmentation challenge dataset (KiTS21) [18]; (b) the liver tumor segmentation benchmark dataset (LiTS) [3]; and (c) the pancreas tumor segmentation branch of the MSD challenge dataset (MSD-Pancreas) [2]. Each dataset is divided into training, validation, and test sets with a ratio of 70%:10%:20%. We use Dice Score and Normalized Surface Dice (NSD) as evaluation metrics to assess both overall accuracy and tumor boundary precision by convention.

Implementation Details. We adopt <u>SAM-B</u> as our image encoder (other SAM models are also used for benchmarking). To ensure consistent intermediate feature dimensions and align textual and visual features, we use CLIP ViT-L/14 as the text encoder. Data augmentation includes random rotations, flips, zooms, and intensity shifts. Foreground and background patches are randomly sampled at a 2:1 ratio during training. All experiments are conducted on a single NVIDIA A100 GPU using AdamW optimizer with the learning rate of $1e^{-4}$ and the batch size of 1 for 200 epochs.

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

We benchmark our **TAGS** method against ten leading 3D medical segmentation techniques, including classical baselines (nnUNet [21], Swin-UNETR [46], UNETR++ [45], 3D UX-Net [26]), 2D SAM-based approaches (SAM-B [24], SAM-Med2D [8]), 3D SAM-based and CLIP-based medical foundation models (SAM-Med3D [50], SegVol [15], and Universal Model [32]), and the original 3D SAM Adapter [16]. For SAM-based models pre-trained on the three evaluation datasets, we perform direct inference on the test sets. By contrast, other models, including **TAGS**, are first trained on the respective training partitions. Table 1 summarizes the comparative results.

Comparison with Classic Benchmarks. TAGS exhibits substantial Dice score gains (exceeding 45% improvement) over both CNN-based [21, 26] and transformer-based [45, 46] methods for kidney and pancreas tumors. In liver tumor segmentation, TAGS still surpasses nnUNet [21] in Dice score, but the margin is smaller, possibly because multiple lesions in the liver require more than a single point prompt. Switching to three prompt points yields a considerable performance boost. TAGS also achieves a 25% NSD advantage over nnUNet, highlighting its effectiveness in resolving complex tumor boundaries.

Comparison with SAM- and CLIP-based Methods. Ta-

Figure 3. Qualitative visualizations of TAGS and SAM-based approaches for kidney, liver, and pancreas tumor segmentation. Lesion areas are highlighted with bounding boxes and zoomed in for detail.

ble 1 and Fig. 3 show that **TAGS** surpasses both slice-by-slice and 3D SAM- and CLIP-based solutions, including SAM-Med3D [50], SegVol [15], and Universal Model [32], despite their pre-training on these target datasets. Although SegVol's "zoom-out-zoom-in" achieves a high Dice for LiTS and Universal Model performs well on MSD, **TAGS** reaches comparable performance with just two extra point prompts. Notably, TAGS maintains superior NSD, confirming its robust boundary delineation.

Comparison with 3D Adapters. **TAGS** outperforms the 3D SAM Adapter [16] by up to 13% in Dice and 8% in NSD, all while retaining a similar parameter count. This gain stems from TAGS's more nuanced capture of small or irregular tumor features (Fig. 3). By embedding both **organ** and **text** prompts, **TAGS** ensures accurate segmentation with fewer interactive prompts, paving the way for broader clinical adoption.

4.3. Ablation Studies

We conduct comprehensive ablation studies to evaluate the design of our framework from various aspects. All experiments use a single-point prompt unless otherwise specified. Table 2 reveals the effectiveness of *organ prompt* and *text prompt alignment*. Extended evaluations are detailed below.

Precision of Organ Prompt. To avoid extra annotation costs, we utilize *TotalSegmentator* [53] to generate organ

Organ Prompt	Text Prompt Alignment	KiTS		LiTS		MSD-Pancreas	
		Dice %	NSD %	Dice %	NSD %	Dice %	NSD %
-	-	76.67	81.95	52.80	69.46	55.02	77.92
✓	-	78.59	85.81	56.67	71.38	55.34	77.78
-	✓	78.56	85.04	57.22	71.42	58.34	80.32
✓	✓	**80.39**	**87.69**	**59.69**	**72.83**	**59.96**	**82.05**

Table 2. Ablation studies of our framework. The best outcomes are highlighted in **bold**.

prompts. Since organ segmentation is relatively easier, *TotalSegmentator* could produce high-quality organ masks even without fine-tuning. *TotalSegmentator* achieves organ segmentation Dice scores of 88.96% on KiTS, 88.57% on LiTS, and 78.54% on MSD-Pancreas, respectively. To examine whether more accurate organ prompts can further improve tumor segmentation performance, we also compare results using ground truth organ prompts for training and testing, as shown in Table 3. The results indicate that tumor segmentation performance could be boosted with more precise organ information.

Effects of Dual-category Text Prompt. Our text prompt design hinges on two key elements: the dual-category foreground and background distinction and the two-level description. Table 6 proves that both are crucial for achieving optimal semantic alignment. For one-category text prompt, we use "{obj} with tumor", where {obj} refers to the organ. For dual-category prompts without two-level descriptions,

Organ Prompt Type	KiTS		LiTS		MSD-Pancreas	
	Dice$_{\%}$	NSD$_{\%}$	Dice$_{\%}$	NSD$_{\%}$	Dice$_{\%}$	NSD$_{\%}$
TotalSegmentator	80.39	87.69	59.69	72.83	59.96	82.05
Ground Truth	83.53	89.36	59.78	75.28	61.32	82.53

Table 3. Comparison results of using totalsegmentator-generated organ mask and ground truth mask.

	KiTS		LiTS		MSD-Pancreas	
	Dice$_{\%}$	NSD$_{\%}$	Dice$_{\%}$	NSD$_{\%}$	Dice$_{\%}$	NSD$_{\%}$
1st	70.51	78.50	49.69	65.79	47.75	71.92
2nd	76.13	84.25	54.32	70.14	54.04	76.97
3rd	78.90	85.98	58.37	72.08	55.89	77.72
4th	80.34	86.59	58.50	72.50	57.34	78.41
Whole Structure	**80.39**	**87.69**	**59.69**	**72.83**	**59.96**	**82.05**

Table 4. Comparison of using aligned feature for segmentation.

we use "healthy {obj}" and "{obj} with tumor" to describe each category. We confirm that including descriptions for both tumor and background improves semantic understanding. Diverse descriptions at the state and template levels also contributes significantly to performance gains.

To further assess the necessity of a medically fine-tuned text encoder, we compare the results of using MedCLIP [52] and our original CLIP text encoder in Table 5. Both approaches perform similarly, though our original setup shows a slight advantage. We attribute this to the text prompt design, which avoids extensive medical priors (e.g., tumor texture, contours) and emphasizes clarity and generality in state-level and template-level descriptions. This design makes the prompts closer to natural language, enabling effective multimodal alignment without fine-tuning the text encoder on medical data. It offers a new way to integrate multimodal information for medical image analysis.

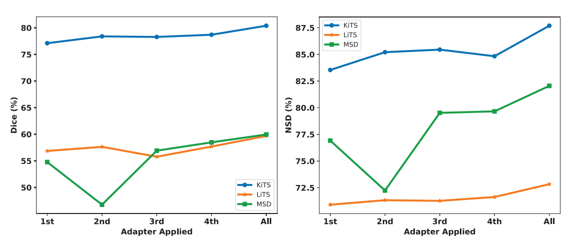

Figure 4. Effectiveness analysis of the necessity of using multi-level alignment adapters. When multi-level adaptation is not used, only one of the four adapters is employed.

Effects of Multi-level Alignment. We first evaluate the necessity of using multi-level alignment adapters. In Fig. 4, we display the segmentation performance when training with each alignment adapter A_s individually. Visualization results for each model on the MSD-Pancreas dataset are shown in Fig. 5. Among the single adapter approaches, models with alignment adapters placed in the third and

Text Encoder	KiTS		LiTS		MSD-Pancreas	
	Dice$_{\%}$	NSD$_{\%}$	Dice$_{\%}$	NSD$_{\%}$	Dice$_{\%}$	NSD$_{\%}$
CLIP VLT-L/14	80.39	87.69	59.69	72.83	59.96	82.05
MedCLIP	79.15	86.23	57.66	72.87	56.70	76.96

Table 5. Ablation results between using original CLIP text encoder and medical-specific MedCLIP text encoder.

Text Prompt Type	KiTS		LiTS		MSD-Pancreas	
	Dice$_{\%}$	NSD$_{\%}$	Dice$_{\%}$	NSD$_{\%}$	Dice$_{\%}$	NSD$_{\%}$
One-category	76.47	83.72	54.59	68.91	52.91	74.89
Dual-category	77.04	84.08	55.82	70.49	54.60	74.94
+state	78.00	84.65	57.62	71.96	57.48	79.14
+template	79.12	86.45	56.66	71.39	56.21	77.22
+both	**80.39**	**87.69**	**59.69**	**72.83**	**59.96**	**82.05**

Table 6. Comparison results of using different text prompts.

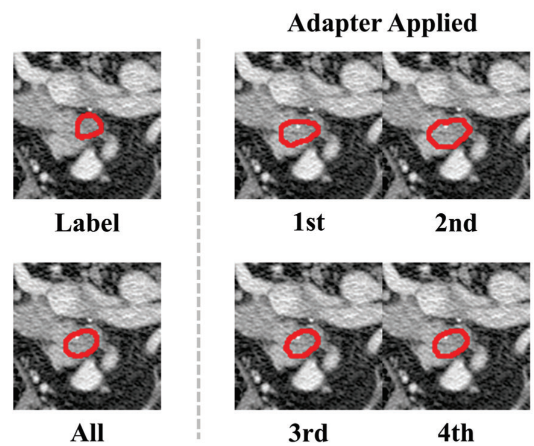

Figure 5. Visualization of segmentation results on the MSD-Pancreas dataset using different alignment strategies.

fourth stages produce results closest to the segmentation ground truth. This aligns with standard foundational model adaptation practices where adapters are positioned near the end of encoders. However, the model using multi-level adapters achieves the highest Dice score and NSD values across all datasets. These results suggest that integrating semantic prompts at multiple stages of the image encoder provides a more refined approach than relying on a single adapter for alignment, allowing the model to learn from textual information more effectively.

We further test the image encoder's ability to learn the spatial features of the image. In this test, we invoke a CLIP-like structure. We remove the prompt encoder and mask decoder while reintroducing the text encoder for feature alignment. Each alignment adapter's output feature $A_s(F_s)$ is aligned with the text feature F_{text}. Linear interpolation is then applied to match input dimensions, producing the segmentation output. We calculate the aligned prediction results for each A_s separately and summarize these results alongside those from the full model structure in Table 4. We observed that as the model progresses through layers, the aligned prediction accuracy produced by adapters at different stages gradually improves. After aligned with the text encoder, predictions from the fourth alignment adapter show an average Dice score difference of 2.41% and an NSD difference of 2.12% compared to the full encoder-decoder model across three datasets. This reveals that the organ and text prompts have enabled the image encoder to learn the image information effectively. Yet, high-quality point prompts can further enhance segmentation accuracy.

Selection Strategy	KiTS		LiTS		MSD-Pancreas	
	Dice %	NSD %	Dice %	NSD %	Dice %	NSD %
random (1pts)	80.39	87.69	59.69	72.83	59.96	82.05
edge (1pts)	78.91	87.39	59.11	72.09	58.02	79.87
edge (3pts)	82.62	90.27	66.47	78.62	58.92	81.12
central (1pts)	82.68	90.29	67.09	81.77	59.05	81.22
ICC (%)	90.54	91.74	90.40	93.81	96.78	96.93

Table 7. Tumor segmentation performance with different point prompt selection strategies. Intra-class Correlation Coefficient (ICC) of each strategy is reported in the last row.

Robustness Against Point Prompts. To explore the robustness of TAGS under different point prompts, we examine how point prompts from various locations impact segmentation performance. In our original training setup, TAGS randomly selects coordinate points from the tumor region as point prompts during inference. Here, we compare the effects of two new point selection strategies on model performance: central selection and edge selection, as illustrated in Fig. 6. central selection uses the central point of the tumor area as the prompt, while edge selection randomly chooses points along the tumor's boundary. For samples with multiple tumors, these operations are applied to the largest tumor only. Table 7 presents the segmentation performance for each strategy, with edge selection tested using one-point and three-point scenarios. We find that both the quality and quantity of point prompts will impact model predictions. When prompt quality is low, such as providing only a single point on the edge, predictions for the tumor region can be more susceptible to interference. Increasing the number of edge prompts or using high-quality central point prompts leads to more accurate predictions.

Interestingly, different point prompts yield only about a 3% variation in segmentation outcomes, thanks to the precision of the proposed adapter. This indicates that, unlike other SAM-based methods, TAGS does not rely heavily on interactive prompts to achieve high-quality predictions. Might be an exception, on the LiTS dataset, using the tumor center point prompt and multiple edge prompts lead to approximately a 10% improvement in both Dice score and NSD. We attribute this to the fact that these combined strategies allow the cropped inference patch around the prompt center to cover more of the tumor region. To further validate the robustness of TAGS with different point prompts, we also calculate the Intra-class Correlation Coefficient (ICC) for the predictions of the above strategies, as shown in Table 7. The results confirm a high consistency in model predictions across different point prompts, with ICC values for Dice score and NSD exceeding 90% in each dataset.

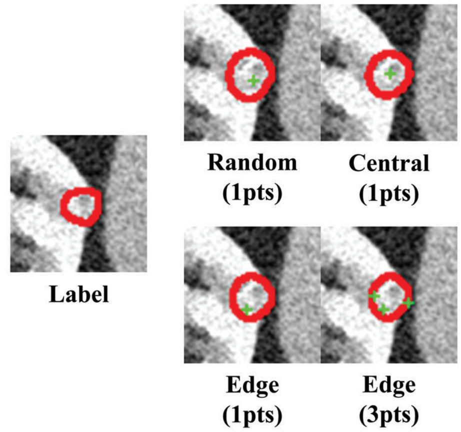

Figure 6. Illustration of different point prompt selection strategies. Green marks on each prediction indicate the selected points.

5. Conclusion

Volumetric tumor segmentation remains a critical challenge in medical imaging due to the limitations of 2D foundation models like SAM in handling 3D spatial relationships and domain shifts. We introduce **TAGS**, a framework that bridges SAM's 2D priors to 3D medical tasks through multi-prompt adaptation, achieving three key advancements: (1) annotation-free organ prompting that leverages automated segmentation tools to guide tumor localization without costly annotations, (2) hierarchical vision-language alignment that integrates CLIP's semantic priors with SAM's spatial acuity, and (3) parameter-efficient 3D adaptation via lightweight multi-stage adapters. Extensive evaluations on three public datasets (kidney, liver, and pancreatic tumors) confirm that **TAGS** substantially outperforms both SAM-based baselines and state-of-the-art 3D segmentation methods, including 3D foundation models specifically trained on medical data. Notably, TAGS delivers these gains with minimal additional parameters, offering a lightweight yet powerful solution for 2D-to-3D model adaptation. Future work could explore extending this multi-prompt strategy to other clinical tasks and integrating additional semantic cues, further advancing the domain-specific capabilities of foundation models in medical imaging. Code and models are released to foster reproducibility.

6. Acknowledgement

This research was supported by NIH under grant numbers R01-HL171376 and U01-CA268808.

References

[1] Rabin Adhikari, Manish Dhakal, Safal Thapaliya, Kanchan Poudel, Prasiddha Bhandari, and Bishesh Khanal. Synthetic boost: Leveraging synthetic data for enhanced vision-language segmentation in echocardiography. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 89–99. Springer, 2023. 3

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 5

[3] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 5

[4] Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. Sam3d: Segment anything model in volumetric medical images. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2024. 2

[5] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 4

[6] Cheng Chen, Juzheng Miao, Dufan Wu, Aoxiao Zhong, Zhiling Yan, Sekeun Kim, Jiang Hu, Zhengliang Liu, Lichao Sun, Xiang Li, et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Medical Image Analysis*, 98:103310, 2024. 2, 3

[7] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. 2

[8] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 5

[9] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 2

[10] Sanghyuk Roy Choi, Jungro Lee, and Minhyeok Lee. Orgunetr: Utilizing organ information and squeeze and excitation block for improved tumor segmentation. *IEEE Access*, 2024. 4

[11] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2

[12] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3956–3960, 2021. 3

[13] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023. 1

[14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[15] Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *arXiv preprint arXiv:2311.13385*, 2023. 2, 5, 6

[16] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, 2024. 2, 3, 4, 5, 6

[17] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 2

[18] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. 5

[19] Chuanfei Hu, Tianyi Xia, Shenghong Ju, and Xinde Li. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. *arXiv preprint arXiv:2304.08506*, 2023. 1

[20] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11375–11385, 2024. 4

[21] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 2, 5

[22] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR)*, pages 19606–19616, 2023. 4

[23] Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Deborah Blacker, et al. Mri-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage*, 46(1):177–192, 2009. 1

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 3, 5

[25] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-sam: Bridging text and image towards universal medical image segmentation. *arXiv preprint arXiv:2403.20253*, 2024. 2

[26] Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076*, 2022. 5

[27] Wenhui Lei, Wei Xu, Kang Li, Xiaofan Zhang, and Shaoting Zhang. Medlsam: Localize and segment anything model for 3d ct images. *Medical Image Analysis*, page 103370, 2024. 2

[28] Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *arXiv preprint arXiv:2401.12665*, 2024. 3

[29] Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *arXiv preprint arXiv:2401.12665*, 2024. 2

[30] Sirui Li, Li Lin, Yijin Huang, Pujin Cheng, and Xiaoying Tang. Text-guided foundation model adaptation for long-tailed medical image classification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 1

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[32] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 3, 5, 6

[33] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyou You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2

[34] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921, 2023. 1

[35] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 2

[36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 1

[37] Fanny Orlhac, Jakoba J Eertink, Anne-Ségolène Cottereau, Josée M Zijlstra, Catherine Thieblemont, Michel Meignan, Ronald Boellaard, and Irène Buvat. A guide to combat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine*, 63(2):172–179, 2022. 1

[38] Linkai Peng, Zheyuan Zhang, Gorkem Durak, Frank H Miller, Alpay Medetalibeyoglu, Michael B Wallace, and Ulas Bagci. Optimizing synthetic data for enhanced pancreatic tumor segmentation. In *International Workshop on Personalized Incremental Learning in Medicine*, pages 35–44. Springer, 2024. 1

[39] Kanchan Poudel, Manish Dhakal, Prasiddha Bhandari, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Exploring transfer learning in medical image segmentation using vision-language models. *arXiv preprint arXiv:2308.07706*, 2023. 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[41] 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

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[44] Sourya Sengupta, Satrajit Chakrabarty, and Ravi Soni. Is sam 2 better than sam in medical image segmentation? *arXiv preprint arXiv:2408.04212*, 2024. 2

[45] Abdelrahman M Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024. 5

[46] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022. 5

[47] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

[48] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[49] Haoyu Wang, Hongrui Yi, Jie Liu, and Lixu Gu. Integrated treatment planning in percutaneous microwave ablation of lung tumors. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4974–4977. IEEE, 2022. 1

[50] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. Sam-med3d: Towards general-purpose segmentation models for volumetric medical images, 2024. 2, 5, 6

[51] Xiaofei Wang, Lai Jiang, Liu Li, Mai Xu, Xin Deng, Lisong Dai, Xiangyang Xu, Tianyi Li, Yichen Guo, Zulin Wang, et al. Joint learning of 3d lesion segmentation and classification for explainable covid-19 diagnosis. *IEEE transactions on medical imaging*, 40(9):2463–2476, 2021. 1

[52] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022. 3, 7

[53] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023. 4, 6

[54] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2, 3

[55] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 171–180. Springer, 2021. 2

[56] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 508–518. Springer, 2023. 4

[57] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[58] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 3

[59] Zheyuan Zhang, Elif Keles, Gorkem Durak, Yavuz Taktak, Onkar Susladkar, Vandan Gorade, Debesh Jha, Asli C Ormeci, Alpay Medetalibeyoglu, Lanhong Yao, et al. Large-scale multi-center ct and mri segmentation of pancreas with deep learning. *arXiv preprint arXiv:2405.12367*, 2024. 1

[60] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021. 4

[61] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 3

[62] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 4