

# TAGS: 3D Tumor-Adaptive Guidance for SAM

## Supplementary Material

### 1. Dataset Details

Table 8 presents detailed information on each dataset. The KiTS dataset [20] originates from the MICCAI 2021 Kidney and Kidney Tumor Segmentation Challenge. The LiTS dataset [3] comes from the MICCAI 2017 Liver Tumor Segmentation Challenge, while the MSD-Pancreas dataset [2] is part of the 2018 Medical Segmentation Decathlon’s pancreas branch. Although all datasets initially included training and test splits, we re-divided the training sets into training, validation, and test subsets due to the lack of organ and tumor annotations in the original test sets. In both training and inference, we used only tumor labels and excluded organ annotations from the process.

### 2. Supplementary Implementation Details

**Data Processing.** Our pre-processing pipeline follows the approach in [17]. We resample anisotropic images to the target spacing, followed by intensity clipping and normalization. For data augmentation, each sample has a 50% probability of undergoing random flipping, rotation, or intensity shifting, and a 30% probability of random zooming. Other settings have been detailed in Sec. 4.1.

**Text Prompt Design.** As described in Sec. 3.3, our text prompt incorporates multiple state-level and template-level descriptions for each category. Fig. 8 exhibits all prompts we employ at both levels, where {obj} refers to specific organ name, and {c} represents a state-level prompt. At state-level, we describe the tumor or background region using varied language. While at template-level, we employ clear, general sentences that include descriptions reflecting the data augmentation conditions for each sample. For each category, the embeddings generated from all descriptions by the CLIP text encoder are averaged to form a final embedding, guiding the image encoder’s understanding of image features. This approach effectively avoids potential biases from using a single text description.

**$\lambda$  Design.**  $\lambda$  is set to 0.2, as it achieves optimal performance in our experiments. We fix it to avoid extra learnable parameters and ensure model efficiency.

### 3. Additional Qualitative Evaluations

**Qualitative Visualization on Other Benchmarks.** As a supplement to Fig. 3, we present a qualitative visualization comparison of TAGS with four classic volumetric segmentation benchmarks and CLIP-based benchmarks in Fig. 9. Both Fig. 3 and Fig. 9 visualize the middle slice within the non-zero pixel range of each volume along the depth dimen-

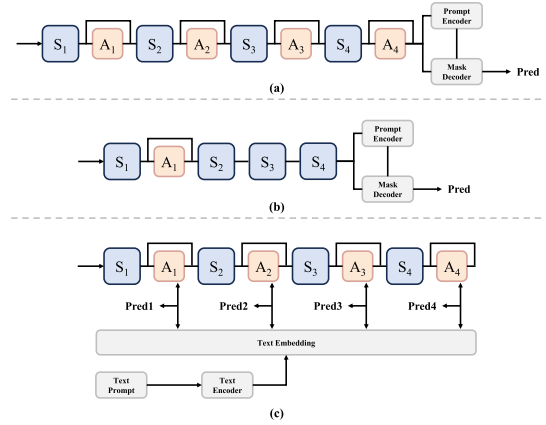


Figure 7. Illustration of different model structures: (a) TAGS when conducting inference; (b) single alignment adapter ablation; (c) directly utilizing aligned feature for prediction.

sion.

It can be seen that, although these benchmarks have succeeded in organ segmentation, their performance in tumor segmentation remains limited. All five models struggle to accurately capture tumor boundaries and shape information. For small pancreas tumors with high variability in shape and texture, nnUNet [23] and 3D UX-Net [28] even show segmentation failures.

**Comparison with Fine-tuned SAM-based Models.** For pretrained SAM-based medical frameworks, though we think the direct inference is a fair comparison (Sec. 4.2), we fine-tuned them on our training sets for further evaluation. As shown in Table 9, each one consistently underperformed relative to TAGS. We also observe a performance decline in some methods after fine-tuning, which is likely due to the limited training set size being insufficient for effective fine-tuning.

### 4. Additional Ablation Studies

**Model Structure of Multi-level Alignment Ablation.** To clarify the distinctions in Sec. 4.3, we present visual representations of the model structures in Fig. 7 for the “single alignment adapter ablation experiment” referenced in Fig. 4, the “CLIP-like structure that directly predicts using aligned features” detailed in Table 4, and the standard TAGS inference model structure.

**Utilizing Different Text Encoders.** Table 10 presents an evaluation of the effectiveness of different text encoders within the TAGS structure. We compare the original TAGS

Dataset	Train	Validation	Test	Resampled Spacing	Intensity Clipping Range	Patch Size	Annotations Included
KiTS	209	30	61	(1, 1, 1)	[-52, 247]	128 × 128 × 128	Kidney, Kidney Tumor, Kidney Cyst
LiTS	83	11	24	(1, 1, 1)	[-17, 201]	128 × 128 × 128	Liver, Liver Tumor
MSD-Pancreas	196	28	57	(1, 1, 1)	[-39, 204]	128 × 128 × 128	Pancreas, Pancreas Tumor

Table 8. Details of the three evaluation datasets.

Category	Method	Kidney Tumor		Liver Tumor		Pancreas Tumor	
		Dice (%)	NSD (%)	Dice (%)	NSD (%)	Dice (%)	NSD (%)
2D SAM-based	SAM-Med2D [8]	0.67	15.75	0.11	2.98	1.47	14.61
3D SAM-based	SAM-Med3D [52]	67.84	81.33	36.33	51.34	60.39	80.64
	SAM-Med3D Turbo [52]	69.16	85.60	38.76	52.45	<u>60.65</u>	80.56
	SegVol [16]	36.30	37.77	23.82	19.28	10.78	17.30
	SegVol w zoom [16]	52.20	50.82	54.27	49.03	26.23	34.30
2D-to-3D	<b>TAGS (1pts)</b>	<u>80.39</u>	<u>87.69</u>	<u>59.69</u>	<u>72.83</u>	59.96	82.05
	<b>TAGS (3pts)</b>	<b>80.83</b>	<b>88.26</b>	<b>66.23</b>	<b>79.33</b>	<b>61.04</b>	<b>83.10</b>

Table 9. The comparison experiments between TAGS and fine-tuned SAM-based benchmarks. The best results are **bold** and the second best ones are underlined.

Text Encoder	KiTS		LiTS		MSD-Pancreas	
	Dice %	NSD %	Dice %	NSD %	Dice %	NSD %
BERT [14]	78.45	85.97	57.22	70.71	46.11	68.90
MedCLIP [54]	79.15	86.23	57.66	<b>72.87</b>	56.70	76.96
CT-CLIP [18]	80.24	86.82	59.21	73.23	55.72	77.86
<b>TAGS</b>	<b>80.39</b>	<b>87.69</b>	<b>59.69</b>	72.83	<b>59.96</b>	<b>82.05</b>

Table 10. Comparison of different text encoders. All experiments are conducted using a single-point prompt, with the best results highlighted in **bold**.

configuration, which uses CLIP, against alternatives such as BERT [14], MedCLIP [54], and CT-CLIP [18] as text encoders. The results indicate that substituting CLIP with BERT-like text-only encoders leads to a decline in performance, falling below the no-text-embedding baseline shown in Table 2. Furthermore, replacing the text encoder with MedCLIP or CT-CLIP produces comparable results, with a performance difference of approximately 3%, reinforcing our assertion that medical fine-tuning was unnecessary for this study.

**Additional Results for Robustness Against Point Prompts.** To further support the content in Table 7, we examine the performance differences between “random selection” and “edge selection” as point prompts increase, as detailed in Fig. 10. For “central selection”, since the prompt point is always at the tumor center, increasing the number of prompts does not introduce new points. Therefore, it is not included in the discussion. We evaluated scenarios with 1, 3, 5, 7, and 10 point prompts. Results show that for “random selection”, increasing the number of points has little impact on performance. Meanwhile, for “edge selection”,

segmentation accuracy stabilizes after 3 prompts, with minimal improvement from more points. This suggests that TAGS achieves strong performance with few prompts, underscoring its efficiency in learning image information and robustness against different point prompts.

**(a) state-level (foreground)**

- c: diseased {obj}
- c: anomalous {obj}
- c: {obj} with flaw
- c: {obj} with illness
- c: {obj} with lesion
- c: {obj} with tumor

**(b) state-level (background)**

- c: {obj}
- c: healthy {obj}
- c: flawless {obj}
- c: perfect {obj}
- c: {obj} without flaw
- c: {obj} without illness
- c: {obj} without lesion
- c: {obj} without tumor

**(c) template-level**

- This is one / a / the {c} in the scene
- There is a / the {c} in the scene
- A 3D photo of one / a / the {c}
- A 3D photo of a / the large {c}
- A 3D photo of a / the small {c}
- A 3D photo of a / the rotated {c}
- A flipped 3D photo of a / the {c}
- A bright 3D photo of a / the {c}
- A dark 3D photo of a / the {c}

(cont'd)

- A good 3D photo of a / the {c}
- A bad 3D photo of a / the {c}
- A corrupted 3D photo of a / the {c}
- A blurry 3D photo of a / the {c}
- A low-resolution 3D photo of a / the {c}
- A close-up 3D photo of a / the {c}
- A cropped 3D photo of a / the {c}

Figure 8. Lists of two-level text prompt description that we used for feature alignment. {obj} denotes the name of specific organ and {c} refers to a state-level description.

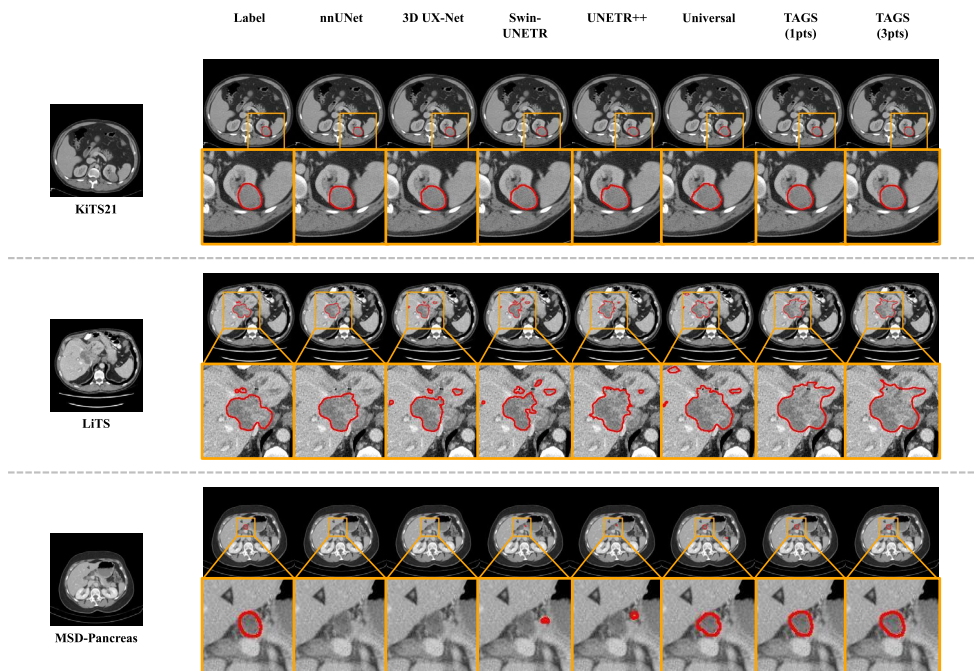


Figure 9. Qualitative visualizations of TAGS and other volumetric segmentation benchmarks approaches for kidney, liver, and pancreas tumor segmentation. Lesion areas are highlighted with bounding boxes and zoomed in for detail.

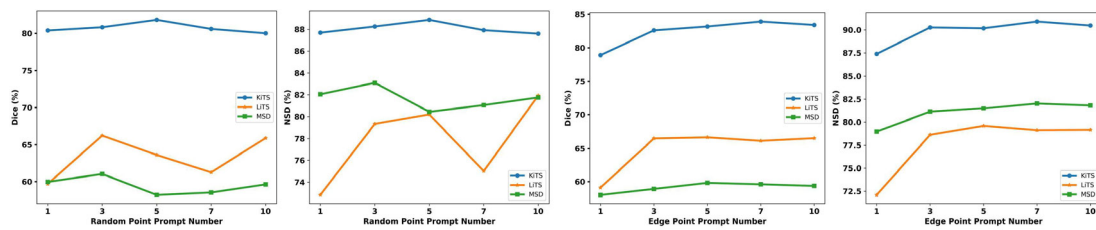


Figure 10. Performance differences between “random selection” and “edge selection” as the number of point prompts increase.