Advancing Generalizable Tumor Segmentation with Anomaly-Aware Open-Vocabulary Attention Maps and Frozen Foundation Diffusion Models

Supplementary Material

A. Dataset Details

Our study utilizes datasets encompassing tumors across 7 diseases and 6 organs, derived from both public and private sources. We summarize all the datasets in Table 7.

A.1. Public Datasets

KiTS23. This dataset is from the Kidney and Kidney Tumor Segmentation Challenge [18], which provides 489 cases of data with annotations for the segmentation of kidneys, renal tumors, and cysts.

MSD. The datasets of liver tumor, pancreas tumor, colon tumor, lung tumor, and brain tumor are part of the Medical Segmentation Decathlon (MSD) [3], providing annotated datasets for various tumors.

BraTS23. This dataset is part of the RSNA-ASNR-MICCAI BraTS 2023 Challenge [1], comprising $1,251$ multi-institutional, clinically-acquired multi-parametric MRI (mpMRI) scans of glioma. The ground truth annotations include sub-regions used for evaluating the 'enhancing tumor' (ET), 'non-enhancing tumor core' (NETC), and 'surrounding non-enhancing FLAIR hyperintensity' (SNFH). In this study, we adopt the 'whole tumor' setting, which describes the complete extent of the disease, for segmentation evaluation.

TotalSegmentator. TotalSegmentator [40] collects 1024 CT scans randomly sampled from PACS over the timespan of the last 10 years. The dataset contains CT images with different sequences (native, arterial, portal venous, late phase, dual-energy), with and without contrast agent, with different bulb voltages, with different slice thicknesses and resolution and with different kernels (soft tissue kernel, bone kernel). A total of 404 patients showed no signs of pathology, and their data are used in our study as healthy samples for anomaly detection training.

A.2. Private Datasets

This dataset comprises a large number of high-resolution T2-weighted 3D MRI images from a total of 400 patients. We acquired one volume from each patient. The segmentation ground truths are provided for each volume in the dataset. All liver tumors and surrounding normal tissues were segmented manually by one radiologist and confirmed by another. During the annotation phase, the radiologists are also provided with the corresponding post-surgery pathological report to narrow down the search area for the tumors. All the MRI scans share the same in-plane dimension of 512×512, and the dimension along the z-axis ranges from 85 to 225, with a median of 155. The in-plane spacing ranges from 0.45×0.45 to 0.62×0.62 mm, with a median of 0.53×0.53 mm, and the z-axis spacing is from 3.0 to 5.5 mm, with a median of 4.2 mm.

A.3. Preprocessing

We adopt similar data processing strategies as used in MAISI [16]. For CT images, the intensities are clipped to a Hounsfield Unit (HU) range of -1000 to 1000 and normalized to a range of $[0, 1]$. For MR images, intensities are normalized such that the 0th to 99.5th percentile values are scaled to the range $[0, 1]$. Intensity augmentations for MR images include random bias field, random Gibbs noise, random contrast adjustment, and random histogram shifts. Both CT and MR images undergo spatial augmentations, such as random flipping, random rotation, random intensity scaling, and random intensity shifting.

B. More Qualitative Analysis.

For qualitative analysis on BraTS23 [1], we present visualizations of segmentation results in Fig. 6. This shows that our approach achieves much better zero-shot cross-modality generalization performance compared with other competing methods.

C. Additional Ablation Experiments

In line with the ablation study setting in the main paper, where the model is trained on the KiTS23 dataset and four CT tumor datasets from MSD, including colon, pancreas, lung, and hepatic vessel tumors, followed by testing on the MSD liver and brain tumor datasets to evaluate generalization to unseen tumors and modalities, we conduct extensive ablation studies for further evaluation.

Significance of Multi-scale Feature Aggregation We aggregate cross-attention matrices between text-attribution keys and pixel queries across three feature levels to generate the AOVA maps. We conduct ablation experiments to examine the efficacy of utilizing multi-scale image features from the MAISI VAE encoder. The outcomes, elucidated Tab. 8, provide a comprehensive understanding of the performance gains achieved through multi-scale feature aggregation for constructing AOVA maps, compared to using single-level image features.

Effectiveness of Latent Space Inpainting. We demonstrate the impact of using versus not using training-free latent space inpainting (LSI) strategy when generating

Data Source	Modality	Dataset Name	Segmentation Targets	Number of scans
Public	CT	KiTS23 [18]	Kidney Tumor, Kidney Cyst	489
		MSD-Colon [3]	Colon Tumor	126
		MSD-Liver [3]	Liver Tumor	131
		MSD-Hepatic Vessel [3]	Hepatic Vessel Tumor	303
		MSD-Lung [3]	Lung Tumor	64
		MSD-Pancreas [3]	Pancreas Tumor	281
		TotalSegmentator [40]	Kidney, Lung, Pancreas, Colon, Liver, Brain	404
	MRI	MSD-Brain [3]	Gliomas	484
		BraTS23 [1]	Gliomas	1251
Private	MRI	in-house dataset	Liver Tumor	400

Table 7. Details of Datasets.

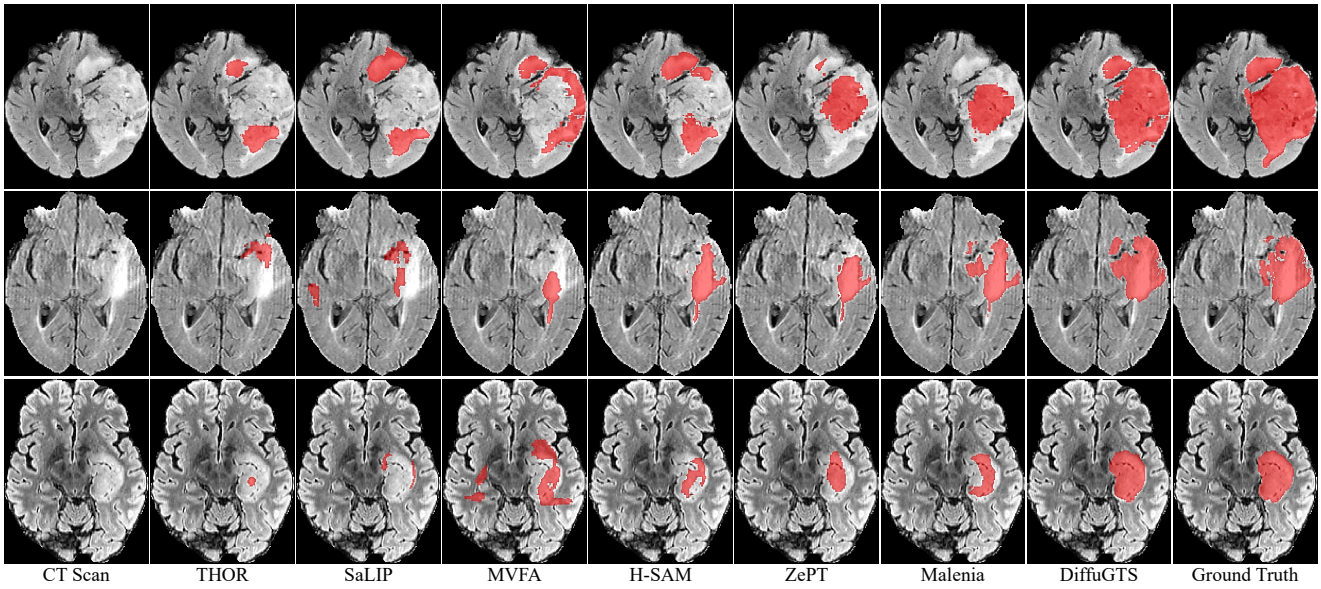

Figure 6. Qualitative visualizations of zero-shot segmentation results on BraTS23 [1].

Feature Levels	MSD Liver Tumor		MSD Brain Tumor	
	DSC↑	NSD↑	DSC↑	NSD↑
Level 1	62.07	72.16	43.40	45.33
Level 2	62.28	72.49	44.92	46.84
Level 3	62.13	72.35	43.88	45.96
Aggregation	**63.23**	**73.58**	**47.51**	**49.75**

Table 8. Ablation study of multi-scale feature aggregation for constructing AOVA maps. The DSC and NSD are reported. The best result is in light blue.

Method	MSD Liver Tumor		MSD Brain Tumor	
	DSC↑	NSD↑	DSC↑	NSD↑
DiffuGTS$_{MAISI}$	60.36	70.31	30.55	32.74
DiffuGTS$_{MAISI}$ + LSI	**63.23**	**73.58**	**47.51**	**49.75**

Table 9. Ablation study on leveraging the latent space inpainting (LSI) strategy to generate pseudo-healthy equivalents, compared to directly using MAISI for generation. The DSC and NSD metrics are reported.

pseudo-healthy equivalents in Tab. 9. Directly applying MAISI for the generation leads to substantial changes in the healthy regions of the target organ (also shown in Fig. 3), which subsequently decreases segmentation performance. In contrast, our strategy effectively preserves details in the organ that are unaffected by the disease, underscoring the importance of modifying the generation process of the original MAISI [16] through latent space inpainting strategy. Additionally, this approach is entirely training-free, avoiding the computational costs associated with retraining or fine-tuning a foundational diffusion model. The illustration of the one-step reverse process of the inpainting strategy is shown in Fig. 7.

Is the improvement solely attributed to the MAISI? To leverage the capabilities of the medical foundational dif-

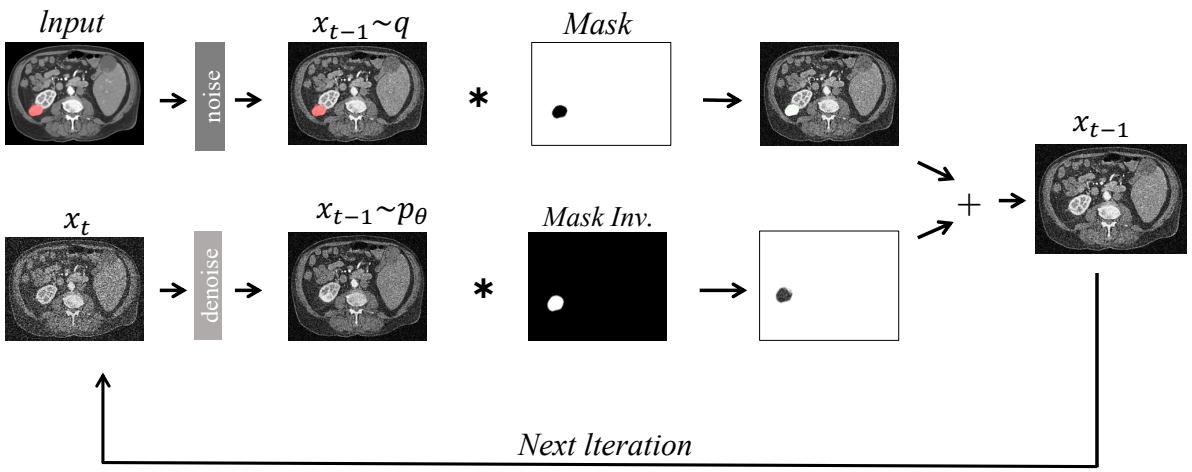

Figure 7. The illustration of the one-step reverse process of the inpainting strategy.

Method	MSD Liver Tumor		MSD Brain Tumor	
	DSC↑	NSD↑	DSC↑	NSD↑
ZePT [22]	59.16	68.72	19.54	22.02
Malenia [23]	59.83	70.08	19.83	22.58
ZePT [22] + MAISI [16]	60.16	70.14	27.21	29.53
Malenia [23] + MAISI [16]	60.28	70.22	27.86	29.94
DiffuGTS	**63.23**	**73.58**	**47.51**	**49.75**

Table 10. Comparisons between **DiffuGTS** and existing methods combined with MAISI.

fusion model, we introduce a series of sophisticated designs and demonstrated their effectiveness through ablation studies. Additionally, we conduct further experiments to show that the performance improvements are not merely due to utilizing the medical foundational diffusion model, but largely stemmed from our innovative designs. To validate this, we apply the MAISI VAE encoder to some existing methods and use MAISI to refine the masks generated by these methods.

The comparison results are shown in Tab. 10. We observe that using the VAE encoder from MAISI for image feature extraction and employing MAISI's generative capability to further refine the masks enhances the performance of existing methods. This supports our motivation for leveraging foundational diffusion models for advanced zero-shot tumor segmentation. Furthermore, even when existing methods benefit from MAISI's capabilities and knowledge, DiffuGTS consistently outperforms them. **This demonstrates that the improvement in zero-shot generalization performance is not solely due to the foundational diffusion model, but also attributed to our innovative designs**, which effectively unleash the potential of utilizing foundation diffusion model for generalizable tumor segmentation.

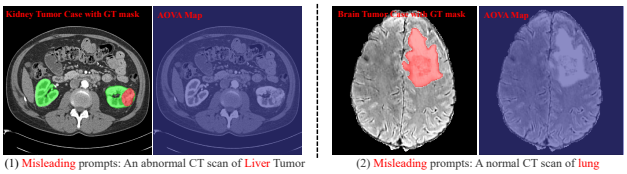

Figure 8. How the model processes misleading text prompts.

D. Model Robustness Analysis

In Fig. 8, we show how our model handles misleading prompts: (1) a disease that is not present, and (2) using a lung-related prompt on a brain scan. The AOVA maps generated by these prompts exhibit no strong activation, indicating that the model recognizes that none of the image content is relevant to the text prompts and therefore does not predict any foreground mask. This further demonstrates that our model has effectively learned the correlations between visual features and textual descriptions, achieving a genuine understanding of anatomical structures.

E. Explanation of "Pseudo-Healthy" Images

We would like to further clarify that the generated pseudo-healthy images are not actual healthy images. Similar to many diffusion-based medical anomaly detection methods [5, 41], the primary purpose of generating these pseudo-healthy images is to segment tumors by highlighting the differences between the original image and the generated image. Ideally, the generated image should exhibit significant changes in the tumor region while preserving the non-tumor areas of the original image, regardless of whether those areas are healthy. Thus, the term "ideal" here specifically refers to tumor segmentation, rather than implying the generation of a completely healthy image. In other words, the

generated pseudo-healthy images only need to preserve the non-tumor areas of the original image while significantly altering the tumor regions, rather than striving to create a fully healthy image. Additionally, whether non-tumor regions of an organ with tumors can still be considered "healthy" is a broader discussion beyond the technical scope of this paper. To prevent any misunderstandings, we refer to these generated images as "pseudo-healthy" images.

F. Analysis of Potential Data Leakage

We used MSD, KiTS23, BraTS23, and an in-house liver tumor dataset for evaluation. Among these, only the MSD overlaps with the dataset used during MAISI's training. A key concern is whether the MAISI framework inadvertently introduces label information leakage that could compromise the model's training independence. In this section, we conduct a rigorous analysis of this critical issue. Apparently, the performance improvement of our framework is not exclusively derived from the MAISI integration. As validated in Tab. 10, the principal performance improvement mainly stems from our innovative designs. Furthermore, we clarify that our framework does not leak any label information from MAISI related to the MSD dataset into downstream testing. First, we use the internal features of the MAISI VAE encoder. The MAISI VAE encoder and decoder were trained on the volume reconstruction task, which only involved image data and did not use any mask annotations. Therefore, using the MAISI VAE encoder's internal features to train the AOVA maps poses no risk of data leakage. Second, the diffusion model in MAISI is trained on the MSD dataset to synthesize tumors explicitly conditioned on a tumor mask via ControlNet. In contrast, our method utilizes a coarse tumor mask implicitly through a repaint mechanism, forcing the model to generate pseudo-healthy organs instead of tumors. This fundamental divergence in conditioning strategies shifts the MAISI's inference paradigm from an in-distribution scenario (tumor generation aligned with MAISI's training data) to an out-of-distribution scenario (synthesizing healthy anatomy from anomalous inputs). This approach essentially prevents the diffusion model from utilizing any memorized label information. If data leakage were to occur, the model would generate the tumor rather than the pseudo-healthy organ we intend. Additionally, generating pseudo-healthy organs on MSD is not involved in MAISI's training. These support the claim that our framework does not leak any label information from MAISI related to the MSD dataset into downstream segmentation testing. Moreover, the superior performance of **DiffuGTS** on KiTS23, BraTS23, and our in-house liver tumor dataset—all excluded from the MAISI foundation model's training data—demonstrates the generalizability and robustness of our proposed strategies.

G. Limitations and Future Work

Our method, through carefully crafted innovative designs, has unleashed the potential of medical foundational diffusion models for advanced zero-shot 3D tumor segmentation. However, it remains constrained by the capabilities of the underlying medical foundational diffusion model. As the MAISI VAE is designed as a foundational model for 3D CT and MRI, our research is similarly limited to these imaging modalities, leaving other modalities, such as 2D X-ray, unaddressed. In future research, we aim to explore zero-shot multimodal models that encompass a broader range of imaging modalities and clinical scenarios. Furthermore, as medical foundational diffusion models continue to evolve, our method stands to benefit from these advancements, with the potential for further enhancement in performance.