ZePT: Zero-Shot Pan-Tumor Segmentation via Query-Disentangling and Self-Prompting

Supplementary Material

6. Appendix

6.1. Dataset Details

Training: In Stage-I, we assemble the training sets of 8 public datasets, including Pancreas-CT [52], AbdomenCT-1K [40], CT-ORG [51], CHAOS [32], AMOS22 [25], BTCV [34], WORD [38] and TotalSegmentator [60]. These datasets exclusively contained organ labels. In Stage-II, we add CT images from the training sets of LiTS [5] and KiTS [20]. The overall seen categories used for training consist of 25 organ classes and 2 tumor classes (Liver Tumor and Kidney Tumor).

(1) Pancreas-CT [52] consists of 82 contrast-enhanced abdominal CT volumes. This dataset only provides the pancreas label annotated by an experienced radiologist, and all CT scans have no pancreatic tumor.

(2) AbdomenCT-1K [40] consists of 1112 CT scans from five datasets and includes annotations for the liver, kidney, spleen, and pancreas.

(3) CT-ORG [51] comprises 140 CT images containing 6 organ classes. This dataset is sourced from eight different medical centers. Predominantly, these images display liver lesions, encompassing both benign and malignant types.

(4) CHAOS [32] provides 40 CT scans including healthy abdomen organs without any pathological abnormalities (tumors, metastasis, and so on) for multi-organ segmentation.

(5) AMOS22 [25], the multi-modality abdominal multi-organ segmentation challenge of 2022, contains 500 CT scans with voxel-level annotations of 15 abdominal organs.

(6) BTCV dataset [34] contains 30 subjects of abdominal CT scans where 13 organs are annotated by interpreters under the supervision of radiologists at Vanderbilt University Medical Center.

(7) WORD [38] collects 150 CT scans from 150 patients before the radiation therapy in a single center. All of them are scanned by a SIEMENS CT scanner without appearance enhancement. Each CT volume consists of 159 to 330 slices of 512×512 pixels. All scans of WORD dataset are exhaustively annotated with 16 anatomical organs.

(8) TotalSegmentator [60] consists of 1024 CT scans of different body parts with a total of 104 labeled anatomical structures. Only organ labels are adopted in this paper.

(9) LiTS [5] contains 131 and 70 contrast-enhanced abdominal CT scans for training and testing, respectively. The data set was acquired by different scanners and protocols at six different clinical sites, with a largely varying in-plane resolution from 0.55 to 1.0 mm and slice spacing from 0.45 to 6.0 mm.

(10) KiTS [20] includes 210 training cases and 90 testing cases with annotations provided by the University of Minnesota Medical Center. Each CT scan has one or more kidney tumors.

Full Name of Tumor Subtype	Count
Adenocarcinoma	278
Mucinous adenocarcinoma	64
Signet ring cell adenocarcinoma (rare)	29
Adenosquamous carcinoma (rare)	17

Table 6. Dataset details of the real-world colon tumor segmentation dataset.

Inference: We employ the MSD dataset [2] that encompasses a range of segmentation tasks for five tumor types in CTs. Among these, pancreas tumors, lung tumors, colon tumors, and hepatic vessel tumors belong to unseen categories. A real-world, private dataset containing 388 3D CT volumes of four distinct colon tumor subtypes is also utilized for testing.

(1) MSD CT Tasks [2] includes liver, lung, pancreas, colon, hepatic vessel, and spleen tasks for a total of 947 CT scans with 4 organs and 5 tumors.

(2) To further evaluate the proposed method, we collect a large real-world colon cancer CT dataset, which consists of 388 patients diagnosed with colon cancer. For each patient, an abdominal CT (venous phase) scan is collected, and the tumor region is annotated by an experienced gastroenterologist and later verified by another senior radiologist. During the annotation phase, the physicians are also provided with the corresponding post-surgery pathological report to narrow down the search area for the tumors. All the scans share the same in-plane dimension of 512×512, and the dimension along the z-axis ranges from 36 to 146, with a median of 91. The in-plane spacing ranges from 0.60×0.60 to 0.98×0.98 mm, with a median of 0.76×0.76 mm, and the z-axis spacing is from 5.0 to 7.5 mm, with a median of 5.0 mm. There are four tumor subtypes in this dataset. The full name and incidence count for each disease are shown in Tab. 6. It is important to note that signet ring cell adenocarcinoma and adenosquamous carcinoma constitute only an exceedingly small proportion of all colon cancer cases, illustrating the long-tailed distribution characteristic of real-world disease incidence.

We summarize all the datasets in Tab. 7. As our main

Datasets	#Target	#Scans	Annotated categories
Pancreas-CT [52]	1	82	Pancreas
AbdomenCT-1K [40]	4	1000	Spleen, Kidney, Liver, Pancreas
CT-ORG [51]	4	140	Lung, Liver, Kidneys and Bladder
CHAOS [32]	4	40	Liver, Left Kidney, Right Kidney, Spleen
AMOS22 [25]	15	500	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, Pan, RAG, LAG, Duo, Bla, Pro/UTE
BTCV [34]	13	30	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, R&SVeins, Pan, RAG, LAG
WORD [38]	16	150	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Pan, RAG, Duo, Col, Int, Rec, Bla, LFH, RFH
TotalSegmentator [60]	104	1024	Spl, RKid, LKid, Gall, Liv, Sto, Pan, RAG, LAG, Eso, Duo, Small Bowel, Colon, and so on.
LiTS [5]	2	201	Liver, *Liver Tumor*
KiTS [20]	2	300	Kidney, *Kidney Tumor*
MSD CT Tasks [2]	9	947	Spl, Liver, *Liver Tumor*, *Lung Tumor*, *Colon Tumor*, Pancreas and *Pancreas Tumor*, Hepatic Vessel and *Hepatic Vessel Tumor*
real-world colon tumor dataset	4	388	*Colon Tumor* with four subtypes

Table 7. The information for all datasets used for training and testing.

objective is not dealing with partial label problem, we directly adopt the successful data processing strategy in [36]. Concretely, we pre-process CT scans using isotropic spacing and uniformed intensity scale to reduce the domain gap among various datasets. Then we unify the label index for all datasets. For these datasets (KiTS, WORD, AbdomenCT-1K, and CT-ORG), which do not distinguish between the left and right organs, we split the organ (Kidney, Adrenal Gland, and Lung) into left part and right part. Since we formulate each organ segmentation result as a binary mask, we can organize the segmentation ground truth for these overlapped organs independently in a binary mask manner. During traning, we associate fundamental queries with organ classes, and advanced queries with tumor classes. The corresponding relationship is shown in Tab. 8.

6.2. Qualitative Analysis on Real-World Colon Tumor Segmentation Dataset.

For qualitative analysis on real-world colon tumors, we present visualizations of segmentation results in Fig. 5. This shows that our approach achieves much better zero-shot segmentation performance on real-world colon tumors compared with other methods.

6.3. Detailed Results of Real-World Colon Tumor Segmentation Analysis.

In Tab. 9, we provide a detailed analysis of the detection and segmentation performance of ZePT and OVSeg [35], the second-ranked method, for four subtypes of colon tumors within the real-world colon tumor dataset. ZePT consistently demonstrates superior detection and segmentation capabilities over OVSeg for both commonly encountered and rare subtypes of colon tumors. In the analysis of segmentation performance for the common colon cancer subtypes, Adenocarcinoma and Mucinous Adenocarcinoma, ZePT exhibited a substantial enhancement over OVSeg [35] in terms of DSC, achieving increases of 33.80% and 22.66%, respectively. In the segmentation of rare colon cancer subtypes, namely Signet Ring Cell Adenocarcinoma and Adenosquamous Carcinoma, ZePT also significantly outperformed OVSeg, achieving DSC improvements of 19.12% and 5.14%, respectively. These results highlight ZePT's superior performance and its promising ability for zero-shot tumor segmentation in real-world settings. Since these rare disease types are individually infrequent, it is impossible to collect them completely. Therefore, we address the thorny problem by exploring and enhancing the model's zero-shot segmentation capability.

6.4. Additional Ablation Experiments

6.4.1 Different Choices of Text Encoder.

We evaluate the performance disparities arising from the use of various models as text encoders for generating text embeddings. In Tab. 10, we observe that employing ClinicalBERT [1] as the text encoder yields a slightly higher Dice Similarity Coefficient (DSC), with improvement of 0.39% compared to the use of the CLIP Text Encoder [50] in the context of the real-world colon tumor dataset. Therefore, we use the ClinicalBERT [1] as the default text encoder setting for ZePT.

6.4.2 Effectiveness of Medical Domain Knowledge.

We examine the performance disparities in generating text embeddings for each organ and tumor when using a conventional prompt [36] (*e.g.*, "a computerized tomography of a [CLS]") versus employing additional domain knowledge automatically derived from the Large Language Model GPT4 [44]. As shown in Tab. 10, incorporating additional domain knowledge results in a notable enhancement in text embedding efficacy, evidenced by an increase of 1.54% in the DSC when using ClinicalBERT [1] as the text encoder, and a 1.51% increase when adopting CLIP Text

Fundamental Queries → Organ Categories	Organ Label Index	Advanced Queries → Tumor Categories	Tumor Label Index
F_1 → Spleen	1	A_1 → Spleen Tumor (unseen)	26
F_2 → Right Kidney	2	A_2 → Kidney Tumor (seen)	27
F_3 → Left Kidney	3	A_3 → Kidney Cyst (unseen)	28
F_4 → Gall Bladder	4	A_4 → Gall Bladder Tumor (unseen)	29
F_5 → Esophagus	5	A_5 → Esophagus Tumor (unseen)	30
F_6 → Liver	6	A_6 → Liver Tumor (seen)	31
F_7 → Stomach	7	A_7 → Stomach Tumor (unseen)	32
F_8 → Aorta	8	A_8 → Aortic Tumor (unseen)	33
F_9 → Postcava	9	A_9 → Postcava Tumor Thrombus (unseen)	34
F_{10} → Portal Vein and Splenic Vein	10	A_{10} → Portal Vein Tumor Thrombus (unseen)	35
F_{11} → Pancreas	11	A_{11} → Pancreas Tumor (unseen)	36
F_{12} → Right Adrenal Gland	12	A_{12} → Adrenal Tumor (unseen)	37
F_{13} → Left Adrenal Gland	13	A_{13} → Adrenal Cyst (unseen)	38
F_{14} → Duodenum	14	A_{14} → Duodenal Tumor (unseen)	39
F_{15} → Hepatic Vessel	15	A_{15} → Hepatic Vessel Tumor (unseen)	40
F_{16} → Right Lung	16	A_{16} → Lung Tumor (unseen)	41
F_{17} → Left Lung	17	A_{17} → Lung Cyst (unseen)	42
F_{18} → Colon	18	A_{18} → Colon Tumor (unseen)	43
F_{19} → Intestine	19	A_{19} → Small Intestinal Neoplasm (unseen)	44
F_{20} → Rectum	20	A_{20} → Rectal Tumor (unseen)	45
F_{21} → Bladder	21		
F_{22} → Prostate	22		
F_{23} → Left Head of Femur	23		
F_{24} → Right Head of Femur	24		
F_{25} → Celiac Trunk	25		

Table 8. The correspondence between object queries and the categories they are responsible for. Only organ categories and seen tumor categories involve voxel-wise annotations. Queries tasked with identifying and segmenting seen organs and tumors receive supervision from both ground truth mask annotations and query-knowledge alignment. Advanced queries responsible for identifying and segmenting unseen tumor categories only have weak supervision from query-knowledge alignment.

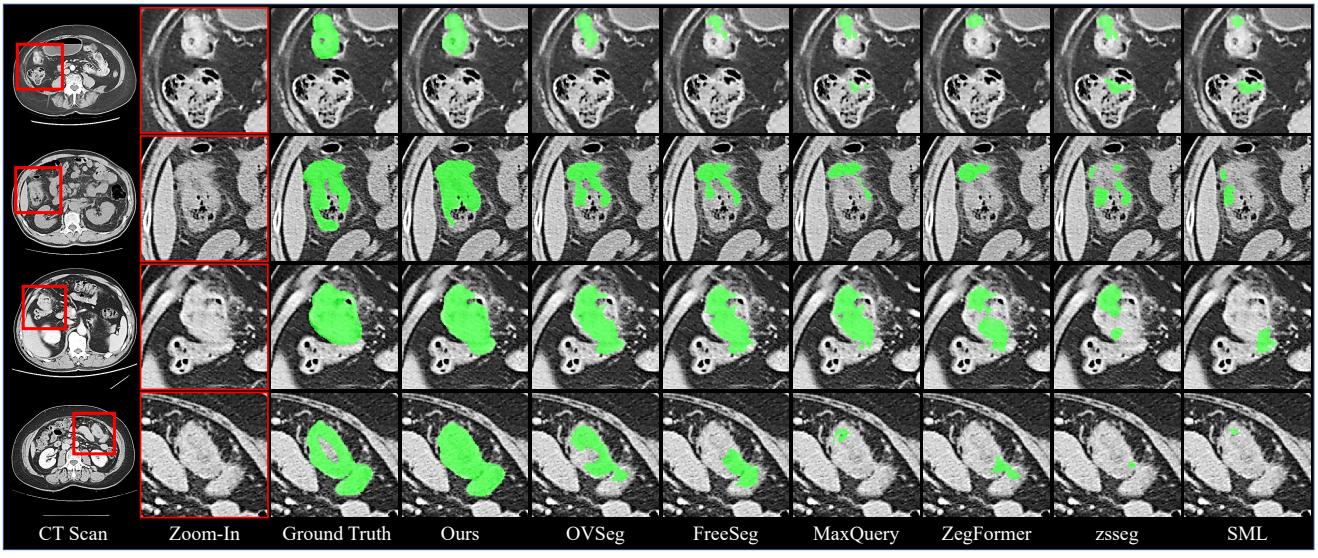

Figure 5. Qualitative visualizations on real-world colon tumor segmentation dataset. We compare ZePT with other advanced OVSS methods and OOD detection methods in a zero-shot manner.

Encoder [50], in the context of the real-world colon tumor dataset. Incorporating medical domain knowledge into the model enhances it with advanced high-level information and detailed visual cues. This improvement boosts the

Method	Adenocarcinoma			Mucinous adenocarcinoma			Signet ring cell adenocarcinoma			Adenosquamous carcinoma			Average		
	AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑
OVSeg [35]	70.06	63.71	16.51	70.01	63.89	16.44	69.93	65.73	15.66	69.80	66.03	15.59	69.95	64.84	16.05
ZePT	**96.44**	**18.58**	**50.31**	**88.29**	**31.64**	**39.10**	**81.80**	**41.05**	**34.78**	**70.87**	**61.89**	**20.73**	**84.35**	**38.29**	**36.23**

Table 9. Detection and segmentation performance of four colon tumor subtypes on real-world colon tumor dataset. We compare ZePT with the second-ranked method OVSeg [35]. ZePT markedly surpasses OVSeg in terms of detection and segmentation efficacy for both prevalent and rare types of colon tumors.

Text Encoder	text sequence	DSC↑
CLIP Text Encoder [50]	A computerized tomography of a [CLS].	34.33
CLIP Text Encoder [50]	A computerized tomography of a [CLS]. + Knowledge	35.84
ClinicalBERT [1]	A computerized tomography of a [CLS].	34.69
ClinicalBERT [1]	A computerized tomography of a [CLS]. + Knowledge	**36.23**

Table 10. Ablation study of the additional medical domain knowledge on the real-world colon tumor segmentation dataset. We also compare the performance disparities in adopting different models as the text encoder (CLIP text encoder vs. ClinicalBERT).

Method	Real-World Colon Tumor dataset		
	AUROC↑	FPR_{95}↓	DSC↑
ZePT-One Stage (No Pretraining)	78.48	44.03	33.19
ZePT-Two Stage (With Pretraining)	**84.35**	**38.29**	**36.23**

Table 11. Ablation study on the impact of pretraining fundamental queries for multi-organ segmentation in Stage-I.

Method	Real-World Colon Tumor dataset		
	AUROC↑	FPR_{95}↓	DSC↑
ZePT (GroupViT [66] backbone)	80.99	41.64	34.47
ZePT (MaskFormer [10] backbone + OFG)	**84.35**	**38.29**	**36.23**

Table 12. Ablation study of the object-aware feature grouping (OFG) strategy and its alternative.

model's discriminative and generalization capabilities. Furthermore, the medical domain knowledge, which is initially auto-generated and then refined by medical professionals, will also be made publicly available alongside the source code.

6.4.3 Importance of Pretraining: One-Stage vs. Two-Stage.

In ZePT, we initially pretrain fundamental queries on datasets exclusively containing organ labels to achieve multi-organ segmentation in Stage-I, subsequently fine-tuning these queries and training advanced queries in Stage-II. However, it is also feasible to bypass Stage-I entirely and directly train the whole model following the training protocol of Stage-II. The performance disparities between one-stage and two-stage training approaches for ZePT are summarized in Tab. 11. ZePT, when trained in two stages, significantly outperforms its one-stage counterpart, demonstrating improved zero-shot colon tumor segmentation. The performance metrics show an absolute increase of at least 5.87% in AUROC, 5.74% in FPR_{95}, and 3.04% in DSC. Furthermore, we observed that the one-stage ZePT variant exhibits significant instability in the initial training phases and necessitates an extended number of epochs for convergence. These phenomena are primarily due to the omission of the initial pre-training stage for fundamental queries, which results in the model's inadequate understanding of anatomical structures, such as organs. As a result, the visual prompts derived from fundamental queries are ineffective in capturing essential information. This inefficacy hinders advanced queries, dependent on the visual prompts, from learning significant features, leading to a compromised feature representation that impairs the model's overall performance. The experimental findings highlight the necessity of a two-stage training approach for the model and reinforce our design insight of commencing with fundamental query training followed by its application in guiding the training of advanced queries.

6.4.4 Object-Aware Feature Grouping vs. other alternatives.

The object-aware feature grouping (OFG) strategy enables object queries in ZePT to acquire organ-level semantics. We compare OFG with a close alternative method, GroupViT [66], which generates a set of queries as clustering centers and clusters pixels with similar semantics via $\mathrm{Gumbel\text{-}Softmax}$ [24, 42] operation. As shown in Tab. 12, adopting OFG results in an improvement of 3.36% in AUROC, 3.35% in FPR_{95}, and 1.76% in DSC, compared to the performance achieved with GroupViT [66]. The results further confirm the efficacy of the OFG strategy. Conversely, the bottom-up clustering approach based on pixel semantics in GroupViT [66] is unsuitable for the zero-shot tumor segmentation (ZSTS) task. This task necessitates differentiating the unseen tumor region from adjacent regions with similar semantics. Therefore, instead of using Gumbel-Softmax for bottom-up pixel grouping, our approach employs it to contrast visual features with specialized object queries. This method prevents the blending of target and adjacent disturbing regions.

6.4.5 Using Different Training Data in Stage-II.

During Stage-II of ZePT's training process, we utilize the LiTS [5] and KiTS [20] datasets, which include two tumor categories, *i.e.*, liver and kidney tumors, respectively. We explore the impact of integrating additional tumor categories into Stage-II of ZePT's training process. We experiment with various tumor categories and corresponding datasets: liver tumor (LiTS [5]), kidney tumor (KiTS [20]), lung tumor (MSD lung task [2]), and pancreas tumor (MSD pancreas task [2]). The efficacy of ZePT, trained across these diverse tumor categories, are evaluated using the MSD hepatic vessel tumor task [2] and the real-world colon tumor dataset for zero-shot tumor segmentation performance. The results are summarized in Tab. 13. We observed several notable intrinsic phenomena.

Firstly, training on images with liver tumors significantly enhances the zero-shot segmentation performance for Hepatic Vessel tumors, more so than training with other tumor categories. This improvement can be attributed to the visual similarity between liver and hepatic vessel tumors in imaging, which results in a substantially higher zero-shot segmentation performance for hepatic vessel tumors following exposure to liver tumors.

Secondly, progressively increasing the number of CT scans and diversifying the tumor categories included in the training process leads to a stable and gradual improvement in the model's zero-shot segmentation performance on unseen tumors.

Thirdly, the model demonstrates significantly enhanced zero-shot segmentation capabilities for unseen tumors that share similar imaging characteristics with the tumor types included in the training set. This is in stark contrast to its performance on tumors with imaging features distinctly different from those in the training set. This phenomenon accounts for the findings in Tab. 1, where the model shows superior zero-shot segmentation for hepatic vessel tumors compared to others, yet demonstrates relatively lower efficacy for lung and colon tumors.

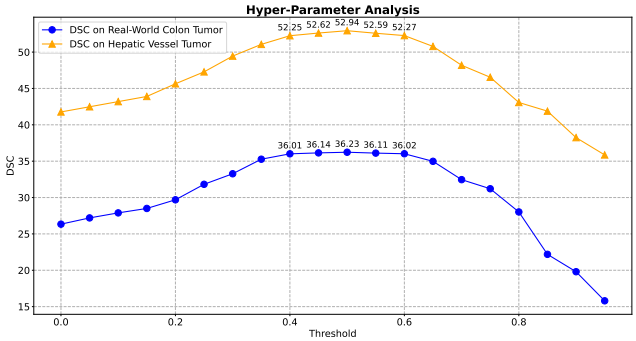

Figure 6. Hyper-Parameter Analysis of the threshold which determines the quality of mask prompts.

6.4.6 Methodological clarification of Gumbel-Softmax.

The main motivation of using Gumbel-Softmax is to make the argmax operation in eq.(4) differentiable [58], where argmax enables the exclusive one-hot hard assignment of each query. This helps queries focus on distinct visual areas without overlap. We conduct experiments to examine the effectiveness of adopting one-hot hard assignment with Gumbel-Softmax. Our findings reveal that, in comparison to the straightforward use of cross-attention, our choice of incorporating one-hot hard assignment alongside Gumbel-Softmax enhances the DSC by 4.95% on the real-world colon tumor dataset. Also, substituting Softmax with Gumbel-Softmax in equation (2) leads to a 2.86% decrease in DSC on the real-world colon tumor dataset. We assume this occurs because the soft assignment by Softmax, prior to the hard assignment of local features, endows the queries with a global receptive field, allowing them to benefit from long-range contexts.

6.5. Hyper-Parameter Analysis

As described in Sec. 3.2, we derive anomaly score maps from the affinity between visual features and fundamental queries. Subsequently, these anomaly score maps undergo min-max normalization, followed by the application of a 0.5 threshold to derive the mask prompts. Then the prompt-based masked attention enables advanced queries to focus on regions specified by the mask prompts, effectively controlling the receptive field of these queries. A critical hyper-parameter in this process, namely the threshold, is instrumental in determining the quality of mask prompts. We conduct experiments on the MSD hepatic vessel tumor task [2] and the real-world colon tumor dataset to analyse the influence of the threshold in Fig. 6. We observed that the model achieves optimal performance in zero-shot tumor segmentation on two datasets when the threshold is set at 0.5. Within the threshold range of 0.4 to 0.6, the model demonstrates high stability and robustness to threshold variations, with no significant changes in performance. However, outside this range, there is a marked decline in performance. This is due to the fact that a threshold near 0 results in minimal masking of visual features, hindering advanced queries from focusing on key visual cues in the lesion area. In contrast, a threshold near 1 leads to excessive masking of visual features, restricting the receptive field of advanced queries and limiting their access to sufficient effective information, consequently causing a marked decrease in performance. Therefore, we adopt 0.5 as the default setting for the threshold.

And another important hyper-parameter is the number of queries. As mentioned in many previous studies [9, 10, 70], the number of queries should be larger than the possible/useful classes in the data, which depends heavily on the data and the task. In our task setup, for the segmentation

Tumor Categories in Dataset	#Scans	MSD [2] Hepatic Vessel Tumor			Real-World Colon Tumor dataset		
		AUROC↑	FPR_{95}↓	DSC↑	AUROC↑	FPR_{95}↓	DSC↑
(Liver Tumor, Kidney Tumor)	341	91.57	20.64	52.94	84.35	38.29	36.23
(Lung Tumor, Kidney Tumor)	274	86.80	33.31	40.57	82.56	40.95	35.02
(Lung Tumor, Liver Tumor)	195	90.72	23.98	50.39	78.50	44.01	33.21
(Pancreas Tumor, Liver Tumor)	413	92.58	19.40	53.83	86.82	34.36	37.95
(Pancreas Tumor, Liver Tumor, Kidney Tumor)	623	93.41	17.79	54.99	87.71	32.29	38.87
(Pancreas Tumor, Liver Tumor, Kidney Tumor, Lung Tumor)	687	94.26	17.05	55.76	89.04	30.68	39.52

Table 13. Ablation study of using different training data in Stage-II.

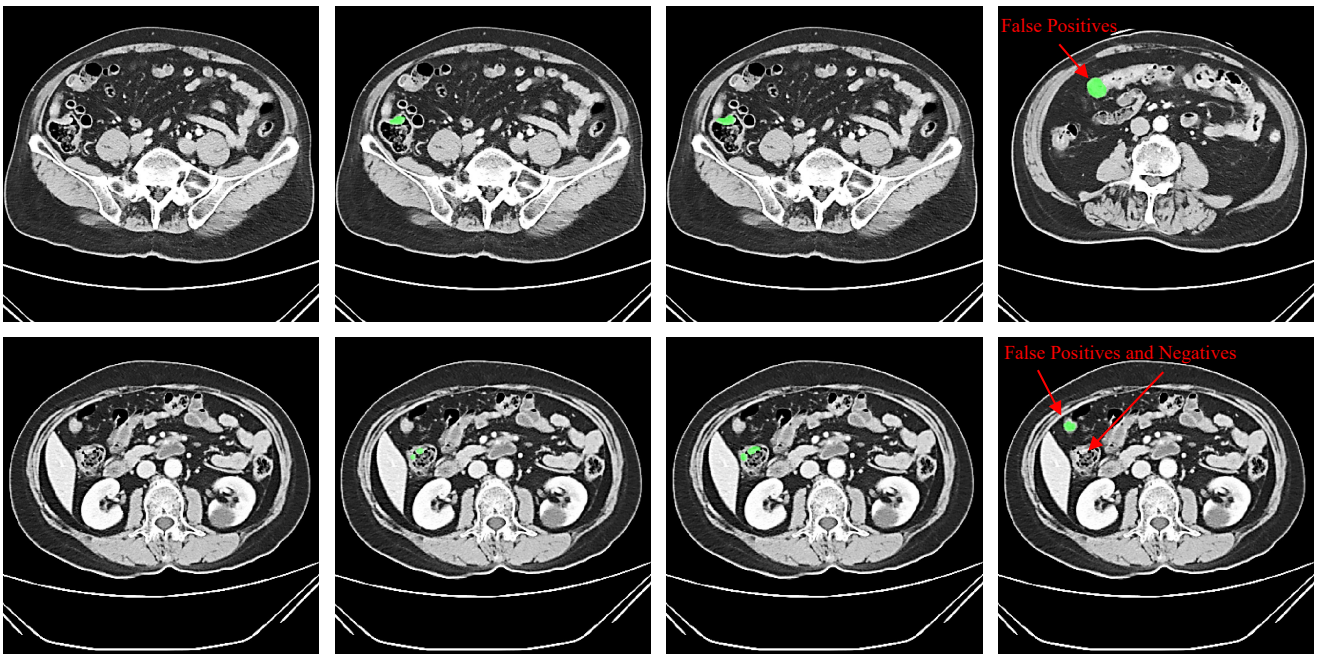

Figure 7. Comparative Visualization: Zero-shot ZePT versus Fully-Supervised nnUNet [23] on real-world colon tumor segmentation dataset. Illustrated are two cases where ZePT successfully detects and segments colon tumors, contrasting with the fully-supervised nnUNet [23], which fails in these instances.

of 25 organs, we assign a fundamental query to each organ. Taking into account the actual occurrences of diseases associated with these organs, we use 20 advanced queries for tumor segmentation. Consequently, in theory, our ZePT model is capable of identifying up to 20 distinct tumor or lesion types. This configuration is flexible and can be modified according to the unique demands of various tasks.

6.6. Comparisons between Different Settings.

Zero-Shot vs. Fully-Supervised. While ZePT demonstrates outstanding performance in zero-shot tumor segmentation, its DSC scores are still lower compared to fully supervised models trained with labels of those unseen tumors. For instance, we trained the robust fully supervised nnUNet [23] model on the collected real-world colon tumor segmentation dataset, achieving a average DSC of 58.30%. This exceeds ZePT's zero-shot colon tumor segmentation performance, which stands at a DSC of 36.23%, by a margin of 22.07 percentage points. Despite this comparison being somewhat unfair, it highlights the substantial room for improvement in zero-shot learning for extremely challenging tasks like tumor segmentation. Notably, ZePT does not fall short in all aspects against the fully supervised nnUNet [23]. Our experiments indicate that the fully supervised nnUNet [23] tends to overfit the training data, resulting in missed detections and false positives in certain cases, particularly with rare tumor types. In contrast, ZePT consistently and accurately identifies and segments colon tumors in these cases. Fig. 7 illustrates two cases in which ZePT successfully segments colon tumors, whereas the fully supervised nnUNet [23] produces false positives and negatives. This comparison highlights the substantial

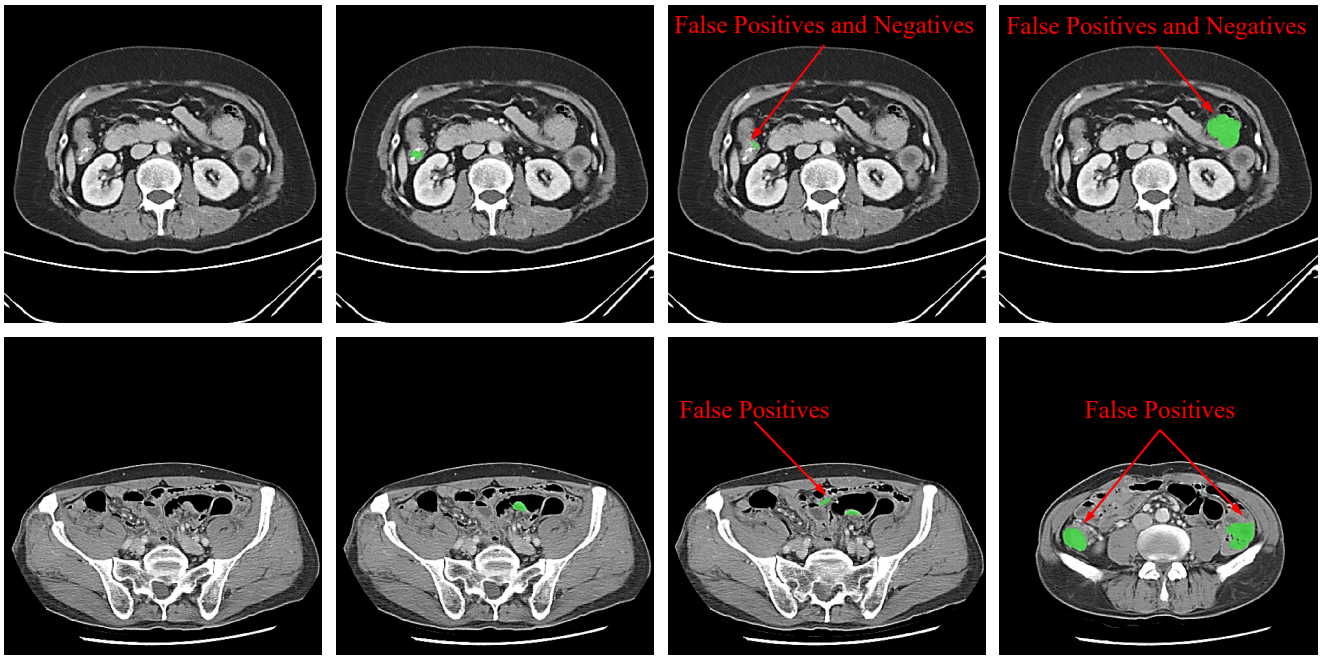

Figure 8. Failure Case Visualizations. Displayed are two examples where both zero-shot ZePT and fully-supervised nnUNet [23] struggle due to the vague and indistinct characteristics of tumor areas.

Method	Task03 Liver		Task06 Lung Tumor	Task07 Pancreas		Task08 Hepatic Vessel		Task09 Spleen	Task10 Colon Tumor
	Organ DSC↑	Tumor DSC↑	DSC↑	Organ DSC↑	Tumor DSC↑	Organ DSC↑	Tumor DSC↑	DSC↑	DSC↑
nnUNet [23]	94.57	58.22	66.57	80.06	50.45	63.29	68.20	96.53	50.07
Swin UNETR [56]	94.14	57.93	68.90	80.18	52.54	62.37	68.63	95.86	50.55
Universal [36]	96.53	71.92	67.11	82.75	60.83	62.64	69.47	96.75	62.15
ZePT (fully-supervised)	**97.22**	**72.95**	**69.07**	**86.23**	**62.10**	**64.39**	**70.65**	**97.04**	**64.87**

Table 14. Benchmark on MSD validation dataset. We compare the fully-supervised ZePT with leading baselines, including nnUNet [23], Swin UNETR [56], and Universal [36] (previously ranked first on the MSD leaderboard), using 5-fold cross-validation on the MSD dataset. The fully-supervised ZePT demonstrated superior segmentation performance overall, particularly in segmenting the pancreas (+3.48%), pancreatic tumors (+1.27%), and colon tumors (+2.72%).

and promising potential of zero-shot learning to address the long-tail distribution challenge in medical imaging.

Fig. 8 displays several failure cases from the real-world colon tumor segmentation dataset, characterized by exceedingly vague and indistinct tumor areas. In these instances, ZePT was unable to detect the tumors. It is important to note, however, that the fully supervised nnUNet [23] model also struggled with these particularly challenging cases.

Fully-Supervised ZePT. Additionally, we trained a fully-supervised version of ZePT to further evaluate its segmentation performance on seen organs and tumors. We follow the settings in [36] and train a strong ZePT model on multiple public datasets. We then conducted a comparative analysis of the fully-supervised ZePT against established baselines, including nnUNet [23], Swin UNETR [56], and Universal [36]. Detailed comparisons based on 5-fold cross-validation on the MSD dataset are presented in Tab. 14. The fully-supervised ZePT achieves overall better segmentation performance and offers substantial improvement in the tasks of segmenting pancreas (+3.48%), pancreatic tumors (+1.27%), and colon tumors (+2.72%). This further demonstrates the novelty and superiority of ZePT's network architecture and training strategy. Notably, ZePT enhances segmentation performance significantly for seen organs and tumors, outperforming previous fully-supervised methods. Furthermore, it exhibits a remarkable ability for zero-shot tumor segmentation, a feature absent in traditional fully-supervised models. These strengths emphasize the importance and potential of ZePT.

6.7. Differences Between ZePT and Existing Zero-Shot Medical Image Segmentation Methods

Research on zero-shot segmentation models is scarcely explored within the medical imaging domain, primarily due

to the intricacies involved in medical image segmentation tasks. Early attempts in this area include [4, 41]. Ma *et al.* [41] proposed a zero-shot CNN model which utilizes two adjacent slices, instead of the target slice, as the input data of deep neural network to predict the brain tumor area in the target slice. This method has the potential to reduce the annotation workload, allowing doctors to only label a subset of the slices. Nevertheless, it requires tumor annotations for training and is capable of segmenting targets only when adjacent slices and their labels are supplied during the training process. Bian *et al.* [4] introduced an annotation-efficient approach based on zero-shot learning for medical image segmentation. This method leverages the information in data of an existing image modality with detailed annotations and transfer the learned semantics to the target segmentation task with a new image modality. Therefore, their approach more closely resembles Domain Adaptation. These existing methods have yet been definitively proven to have the capability to segment multiple tumors in a strictly "zero-shot" manner. To the best of our knowledge, ZePT represents the first method capable of achieving zero-shot pan-tumor segmentation.

6.8. Discussions on Future Works.

Improving Zero-Shot Tumor Segmentation Performance. The analysis of data from Tab. 1 and Tab. 13 indicates that the model excels in zero-shot segmentation for unseen tumor categories that exhibit visual features similar to those of seen tumor categories. This suggests that simulating lesion features akin to unseen tumor categories during training, and directing the model to emphasize these features, could markedly improve its zero-shot segmentation capabilities. Consequently, future studies could investigate the use of diffusion-based models for simulating diverse tumor lesions' visual features and incorporating them into the training regime, which would potentially augment the model's effectiveness in zero-shot tumor segmentation.

Adaptation to Diverse Imaging Modalities. This paper primarily explores the zero-shot tumor segmentation challenge, without delving into addressing the differences between various imaging modalities. Consequently, in line with prior research [8, 36] on creating universal segmentation models for various organs and tumors, our experiments and analyses were solely conducted using CT images. However, our approach is, in theory, adaptable and could potentially be applied to other 3D medical imaging modalities, including MRI and ultrasound. We aim to investigate this potential in future studies.

Expanding Data Collection to Encompass a Broader Spectrum of Tumor Types for Evaluation. In our research, we assembled a dataset of 388 patients with colon tumors and utilized most of the publicly available tumor datasets to develop and evaluate the ZePT model's zero-shot performance. As previously noted, ZePT theoretically possesses the ability to segment a diverse array of unseen tumor types, beyond the scope of currently available datasets. To this end, we are actively compiling a more comprehensive dataset that includes a wider variety of tumor types, aiming to further assess ZePT's capabilities. Although the pervasive issue of data scarcity continues to challenge medical AI model development, ZePT represents a significant stride in overcoming this hurdle through zero-shot learning.