RadOmni: Advancing Foundation Model for Non-contrast CT with Omni Radiology Knowledge

Fenghe Tang[1,2,3*] **Weiwei Cao**[2,3*] **Wenxin Ma**[1] **Kai Cao**[4]
Huanhuan Liu[5] **Ling Zhang**[2] **Shaohua Kevin Zhou**[1†] **Jianpeng Zhang**[2,3†]
[1]Jiangsu Key Laboratory of Multimodal Digital Twin Technology, Suzhou, Jiangsu, China
[2]DAMO Academy, Alibaba Group
[3]Hupan Lab, Hangzhou, China
[4]Department of Radiology, Shanghai Institution of Pancreatic Disease, Shanghai, China
[5]Department of Radiology, Xinhua Hospital,
Shanghai Jiao Tong University School of Medicine, Shanghai, China

Abstract

Non-contrast CT (NCCT) is widely used in clinical practice, yet its limited soft-tissue contrast often renders visual information insufficient for accurate diagnosis. In routine radiology workflows, NCCT is commonly accompanied by contrast-enhanced CT, radiology reports, and pathology reports, which provide complementary knowledge at different levels, from structural and morpho-functional cues to semantic findings and pathological evidence. However, existing pretraining paradigms address this challenge only partially, each introducing complementary but still limited constraints on NCCT representations. To address this limitation, we propose RadOmni, a unified omni-pretraining framework that injects multimodal radiology knowledge into representation learning for non-contrast CT. RadOmni formulates a curriculum consolidation learning strategy, in which knowledge is progressively injected according to optimization difficulty and knowledge level through four stages: masked image modeling on non-contrast CT, cross-phase transfer of anatomical structures and enhanced patterns from contrast-enhanced CT, report-guided semantic learning, and pathology-informed diagnostic enhancement. To mitigate optimization conflicts across heterogeneous objectives and reduce knowledge forgetting, each stage retains and jointly optimizes the objectives from preceding stages, while early encoder layers are selectively frozen to preserve generic representations. To enable such multi-source pretraining, we curate Omni-CT10K, a large-scale CT dataset with more than 10K NCCT scans featuring clinically co-occurring multimodal supervision. Extensive experiments on Omni-CT10K, two in-house benchmarks from different centers, and the public benchmarks show that RadOmni consistently achieves strong performance across classification, segmentation, pathological TNM staging, and radiology report generation tasks. Further analyses demonstrate favorable scaling behavior, robust cross-center generalization, and transferable representations.

1 Introduction

Non-contrast CT (NCCT) characterizes anatomy through intrinsic tissue attenuation differences without the use of exogenous contrast agents [1], making it a safe, low-cost, and widely accessible modality in routine radiology. However, due to its limited soft-tissue contrast, NCCT often provides

*Equal Contribution. This work was done during Fenghe Tang's internship at DAMO Academy.
†Correspondence to: Jianpeng Zhang(jianpeng.zhang0@gmail.com), Shaohua Kevin Zhou (skevinzhou@ustc.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

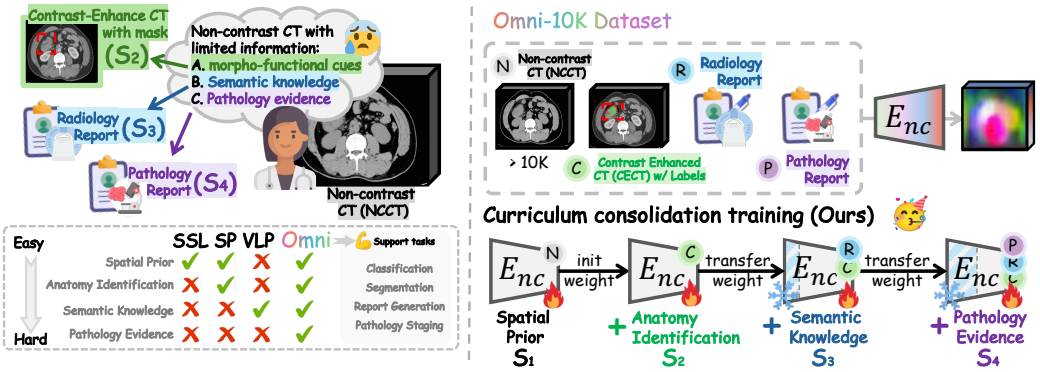


Figure 1: **Overview of the limitation of non-contrast CT (NCCT), previous methods and our RadOmni framework for radiology representation learning, supported by our Omni-CT10K dataset.** SSL: self-supervised learning; SP: supervised pretraining; VLP: vision-language pretraining. **(Left)** NCCT is limited by its non-contrast nature. Unlike SSL, SP, or VLP, RadOmni jointly integrates complementary clinical signals from contrast-enhanced CT, manual label, radiology reports, and pathology reports, enabling the model to learn spatial priors, anatomical semantics, and pathology evidence within a unified framework. **(Right)** Omni-CT10K dataset overview and RadOmni curriculum training introduction. Our curriculum consolidation strategy progressively transfers knowledge from spatial priors to anatomy identification, semantic understanding, and pathology evidence, resulting in a more comprehensive radiology representation encoder.

less clear visual information, making accurate identification challenging for both clinicians and learning models [2, 3, 4]. In clinical practice, the limitation in such subtle visual patterns is often alleviated through additional information acquired along the diagnostic workflow. Clinicians typically assess NCCT together with *contrast-enhanced CT* on *region of interests*, which provides richer morpho-functional cues [5], followed by *radiology reports* that summarize high-level imaging findings [6, 7]. In cases with suspected malignancy or high uncertainty, *pathology reports* may be further obtained to provide definitive diagnostic evidence [8, 9]. Together, these sources support clinical decision-making accurately. However, for Computer-Aided Diagnosis (CAD) models, such heterogeneous information is typically not available during inference, due to factors such as acquisition cost, procedural constraints, and reporting delays[10]. This practical gap motivates an important question: ***how to leverage complementary heterogeneous information during pre-training*** to improve the diagnostic performance of NCCT-based models?

Achieving this challenging goal remains an appealing yet unrealized vision, and existing medical pre-training paradigms provide only partial solutions without fully addressing it. Image-driven self-supervised approaches capture intrinsic structural priors from its visual context, yet lack explicit clinical grounding [11, 12, 13, 14, 15]. Supervised pretraining introduces clinically relevant signals, but remains constrained by predefined annotation schemes and task-specific taxonomies [16, 17, 18]. Cross-modal transfer from contrast-enhanced CT offers additional diagnostic cues, although such strategies are often limited to specific structures or tasks [19, 20]. Vision-language pretraining incorporates radiology reports to provide high-level semantic guidance, yet these signals are often incomplete and loosely grounded, making it difficult to resolve fine-grained visual ambiguity [6, 7, 21, 22, 23, 24]. Taken together, these approaches introduce *complementary but still limited knowledge* on NCCT representations. While they show the promise of heterogeneous supervision, they also expose the need for more principled training strategies that can effectively integrate heterogeneous information and learn general representation beyond their own visual limitations.

To address the above limitation, we first establish a data foundation that exposes heterogeneous yet complementary clinical knowledge for NCCT learning. Specifically, we curate **Omni-CT10K**, a large-scale radiology dataset with clinically co-occurring multimodal supervision, including over 10K NCCT scans with associated radiology reports, pathology reports, and expert-annotated organ and tumor masks. This dataset provides diverse yet complementary clinical signals, enabling the construction of heterogeneous information for NCCT representation learning.

Building upon this data foundation, we propose **RadOmni**, a **Rad**iology **Omni**-pretraining framework for non-contrast CT, aiming to *progressively learn complementary knowledge of NCCT* through heterogeneous supervision. We formulate omni-pretraining as a process of progressive accumulation, where different modalities introduce distinct yet complementary knowledge at different stages. To ensure these knowledge are effectively integrated, we propose a curriculum consolidation learning strategy (Fig. 1 right), which mitigates optimization conflict and prevents prior knowledge forgetting. Concretely, our pretraining follows a four-stage knowledge ladder: *Stage 1* learns structural priors from NCCT, *Stage 2* introduces anatomical and morpho-functional knowledge from contrast-enhanced CT and supervised lesion and organ masks, *Stage 3* incorporates semantic knowledge from radiology reports, and *Stage 4* injects pathology-informed diagnostic knowledge from pathology reports. Crucially, these stages are not optimized in isolation: objectives from earlier stages are preserved, and selective layer freezing stabilizes previously acquired representations, enabling cumulative and consistent knowledge integration. This progressive accumulation steadily enriches the representation space and ultimately enables general capabilities beyond NCCT itself. Extensive experiments on internal, external, and public benchmarks demonstrate that RadOmni achieves strong performance across diverse tasks. In addition, our further analysis shows that RadOmni exhibits favorable data scaling, robust generalization, and transferable representations.

Our key contributions are summarized as follows:

(1) We establish a data foundation for addressing the information-constrained nature of NCCT learning by curating Omni-CT10K, a large-scale radiology dataset with clinically co-occurring multimodal supervision. The dataset exposes heterogeneous yet complementary clinical knowledge, enabling the explicit construction of signals for NCCT representation learning.

(2) We propose RadOmni, a radiology omni-pretraining framework that formulates pretraining as progressive knowledge accumulation under heterogeneous supervision. By cumulatively injecting and consolidating multimodal knowledge, RadOmni effectively expands the representation space of NCCT while mitigating optimization conflict and forgetting.

(3) Through extensive experiments, we show that RadOmni consistently achieves improved performance over existing pretraining paradigms across diverse benchmarks, demonstrating stronger generalization and contributing to more discriminative NCCT representations.

2 RadOmni

2.1 Overview

Problem formulation. As a foundation model, the primary objective is to learn an omni-representative visual encoder for non-contrast CT (NCCT), denoted as $E_{nc} : X_{nc} \longrightarrow Z$, where $Z = \{z_i\}_{i=1}^{n}$ denotes the multi-scale representations. The goal is for Z to capture rich structural, anatomical, and semantic information, enabling robust and generalizable representations for diverse downstream clinical tasks. To achieve this, we leverage multiple complementary sources of supervision from heterogeneous clinical data (e.g., masked image modeling, contrast-enhanced CT, radiology reports, and pathology reports) to progressively enrich the learned representations.

Training with curriculum consolidation. To effectively integrate these heterogeneous supervision signals, we organize omni-pretraining as a curriculum over increasing knowledge level and optimization difficulty. Specifically, the training process is divided into four stages, denoted as S_1, S_2, S_3, S_4, progressing from image-level structure, to anatomy, to report semantics, and finally to pathology-informed diagnostic signals. Each stage introduces a new modality-related learning objective. As training proceeds to later stages, the newly introduced objective is jointly optimized with the retained objectives from earlier stages, while partial layer freezing is applied to preserve generic capabilities. This design allows the encoder to continuously absorb new multimodal knowledge while mitigating forgetting of previously learned capabilities.

Architecture. RadOmni is built around a NCCT encoder E_{nc}, which serves as the core representation backbone throughout pretraining. A lightweight decoder D_{nc} is attached to E_{nc} to support image-level objectives such as masked image modeling and segmentation. To introduce complementary supervision from heterogeneous data, we incorporate several auxiliary modules. Specifically, a pretrained and frozen contrast-enhanced CT network, consisting of encoder E_{ce} and decoder D_{ce}, provides anatomical and morpho-functional guidance. A text encoder E_{text} is used to encode

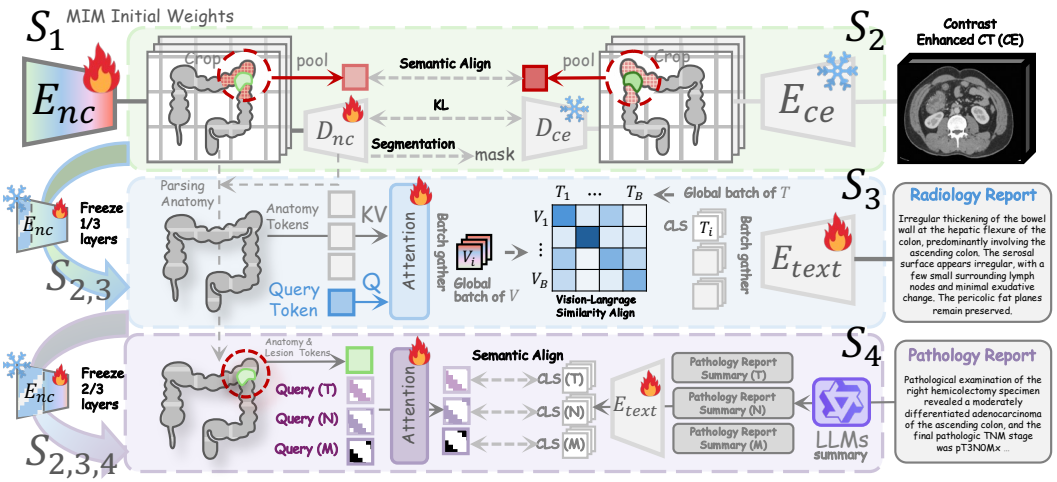


Figure 2: **RadOmni Framework.** The pretraining procedure is divided into four stages: masked image modeling in S_1 for initial weights, anatomical awareness learning from contrast-enhanced CT with segmentation in S_2 (row 1, green), semantic alignment learning between images and radiology reports in S_3 (row 2, blue), and pathology-informed TNM prototype learning from pathology reports in S_4 (row 3, purple). Earlier-stage objectives are retained and jointly optimized in later stages, while early layers are frozen to preserve general capabilities (left arrows indicate progression across stages).

radiology and pathology reports for cross-modal semantic alignment. The overall training follows a staged curriculum, where different modules are progressively activated to introduce increasing levels of supervision. After pretraining, only the NCCT encoder E_{nc} is retained and finetuned for downstream tasks. Details are presented in App. Sec. B.

2.2 S_1: Self-supervised learning for structural priors

The objective of S_1 is to learn low-level structural priors. To this end, we adopt a masked image modeling strategy. Specifically, the input volumes are randomly masked. Following prior approaches [25, 26, 11], sparse operators [11] are applied throughout the multi-stage encoding process to preserve consistency with the masked input positions. Subsequently, learnable tokens are introduced in the decoding stage to reconstruct the original volumes.

2.3 S_2: Cross-phase anatomical and morpho-functional transfer

While S_1 learns generic structural priors from NCCT, its ability to represent tumor and peritumoral morpho-functional patterns remains limited. NCCT is inherently less sensitive in soft-tissue characterization and tumor delineation, whereas these are key strengths of contrast-enhanced CT. The goal of S_2 is therefore to transfer such cross-phase anatomical and morpho-functional knowledge into the encoder E_{nc}. To this end, we initialize E_{nc} from S_1 and introduce a pretrained contrast-enhanced CT model, as a stable teacher for enhancement-aware anatomical representation transfer. Based on this teacher-student setup, we design a cross-phase transfer strategy that promotes both anatomical topological consistency and local prediction consistency.

Cross-phase topological contrast learning. Direct feature matching across phases may emphasize appearance similarity while overlooking the tumor-context relationship that carries the main complementary signal. We therefore adopt an anatomy-guided cross-phase contrastive loss $\mathcal{L}_{cont}$ as the primary alignment objective. For each paired non-contrast and contrast-enhanced CT case, prototype features for the tumor (z_{tum}) and its adjacent contextual organ regions (z_{ctx}) are extracted by masked average pooling using the corresponding masks. We then compute a relative direction vector $d = z_{tum} - z_{ctx}$, which captures the contrast between the tumor and its surrounding anatomy. By aligning these directional representations across phases, the model is encouraged to preserve peritumoral topological consistency:

$$\mathcal{L}_{cont} = 1 - \langle d_{nc}, d_{ce} \rangle , \quad \text{where } d_m = z_{tum}^{(m)} - z_{ctx}^{(m)} . \tag{1}$$

where $\langle a, b \rangle$ refers to the cosine similarity between vectors a and b, more details of implementations are provide in App. Sec. B.

Auxiliary local prediction distillation. To provide complementary local supervision, we further introduce a distillation loss $\mathcal{L}_{KL}$ as an auxiliary regularizer:

$$\mathcal{L}_{KL} = \mathbb{E}\left[D_{KL}(P_{nc} \| P_{ce})\right], \tag{2}$$

where P_{nc} and P_{ce} denote the segmentation predictions from the decoders of the NCCT and contrast-enhanced CT branches, respectively, and D_{KL} denotes the KL divergence. To anchor the transferred knowledge to explicit anatomy, S_2 also optimizes tumor and organ segmentation. The total loss for S_2 is defined as:

$$\mathcal{L}_{S_2} = \mathcal{L}_{seg} + \mathcal{L}_{KL} + \mathcal{L}_{cont}, \tag{3}$$

where $\mathcal{L}_{seg}$ denotes the dice and cross-entropy loss for tumor and organ segmentation. $\mathcal{L}_{KL}$ and $\mathcal{L}_{cont}$ are optimized only when the paired contrast-enhanced CT is available.

2.4 S_3: Report-guided semantic learning

Radiology reports are condensed textual summaries of image semantics written by radiologists. They inherently provide disease-level semantic labels and organ-level findings descriptions. The rich disease and findings semantics in radiology reports can effectively complement NCCT in learning both anatomical and disease-related semantic representations. Therefore, the goal of S_3 is to enhance the semantic perception capability of E_{nc} using radiology reports.

Specifically, on the vision side, we use the organ masks predicted by the decoder D_{nc} to extract multi-scale organ features from the encoder E_{nc}. Learnable visual query tokens are then appended to these extracted features and updated through cross-attention layers. The updated query tokens are further passed through a linear projection layer to generate an anatomy-specific visual embedding V. On the text side, we extract organ-related descriptions from the report with LLMs and encode them into text embeddings T using the text encoder E_{text}. For a batch of visual embeddings $\{V_1, V_2, \ldots, V_B\}$ and their corresponding text embeddings $\{T_1, T_2, \ldots, T_B\}$, we compute the softmax-normalized image-to-text similarities p^{i2t} and text-to-image similarities p^{t2i} as follows:

$$p_i^{i2t} = \frac{e^{\langle V_i, T_i \rangle / \tau}}{\sum_{j=1}^{B} e^{\langle V_i, T_j \rangle / \tau}}, p_i^{t2i} = \frac{e^{\langle T_i, V_i \rangle / \tau}}{\sum_{j=1}^{B} e^{\langle T_i, V_j \rangle / \tau}} \tag{4}$$

where τ is a learnable temperature parameter. The report-guided semantic learning loss is defined as:

$$\mathcal{L}_{semantic} = \frac{1}{2}\left(\frac{1}{B}\sum_{i=1}^{B}(\mathcal{L}_{ce}(y_i^{i2t}, p_i^{i2t}) + \mathcal{L}_{ce}(y_i^{t2i}, p_i^{t2i}))\right) \tag{5}$$

where $\mathcal{L}_{ce}$ is cross-entropy loss. y^{i2t} and y^{t2i} denote one-hot matching labels, where the positive pair is assigned probability 1 and all negative pairs are assigned probability 0. The total loss for S_3 combines S_2 loss with the semantic learning loss:

$$\mathcal{L}_{S_3} = \mathcal{L}_{semantic} + \mathcal{L}_{S_2}, \tag{6}$$

2.5 S_4: pathology-informed diagnostic enhancement

Pathology serves as the clinical gold standard and provides microscopic evidence for the biological behavior of tumors. Connecting CT with pathology allows complementary macroscopic and microscopic cues to enrich the diagnostic representation learned by E_{nc}. In S_4, we introduce pathology-informed cues to further enhance the diagnostic semantics of the encoder. To inject such pathological knowledge, we first employ LLMs to structure pathology reports and extract TNM-related descriptions (in App. Sec. B), including T (tumor invasion), N (node involvement), and M (metastasis). These structured descriptions provide rich pathology-derived diagnostic information, such as invasion depth and lymph node status, which is difficult to infer from CT image pixels alone.

Specifically, to capture pathology-related diagnostic characteristics, we extract visual descriptors of organs and tumors using corresponding masks. We then construct three learnable query tokens

that represent visual prototypes for T, N, and M staging, respectively, and refine them through a cross-attention layer. The refined query tokens are subsequently passed through separate linear projection layers to generate three visual embeddings, $\{V^T, V^N, V^M\}$. For the structured pathology descriptions, we extract the corresponding text embeddings $\{T^T, T^N, T^M\}$ using the text encoder. Finally, we align each visual prototype with its corresponding text embedding. The total loss for S_4 is defined as:

$$\mathcal{L}_{S_4} = \mathcal{L}^T_{semantic} + \mathcal{L}^N_{semantic} + \mathcal{L}^M_{semantic} + \mathcal{L}_{S_3}, \tag{7}$$

where $\mathcal{L}^T_{semantic}$, $\mathcal{L}^N_{semantic}$, and $\mathcal{L}^M_{semantic}$ are optimized only when paired pathology report description is available. These losses share the same formulation as $\mathcal{L}_{semantic}$ in S_3.

3 Omni-CT10K Pretraining Dataset

To support RadOmni pretraining, we curate Omni-CT10K, a dataset comprising 10,591 CT scans from unique patients, along with associated radiology reports, pathology reports, and manually annotated organ and tumor masks provided by senior radiologists. The dataset includes 3,936 abnormal cases and 6,655 normal cases. All tumor cases are confirmed by pathology and/or surgery, while normal cases are verified by colonoscopy. All CT scans are acquired preoperatively, with an interval of no more than three months between imaging and pathology reports. To facilitate multi-source supervision, we employ Qwen3-Max [27] to extract organ-related and TNM-related descriptions from radiology and pathology reports, respectively. The extracted information is used to construct structured supervision signals. A randomly subset is reviewed by individual clinicians to assess the quality of the extracted information. We utilize the entire dataset for pretraining. Additional details on Omni-CT10K are provided in App. Sec. B.

4 Experiment

We conduct extensive experiments to validate the effectiveness of RadOmni. We present our experimental details in Sec. 4.1. Downstream experiments are presented in Sec. 4.2. Ablation study is presented in Sec. 4.3. Data scaling analysis is presented in Sec. 4.4.

4.1 Experimental setup

Finetuning and evaluation benchmarks. To ensure strict center-level separation, the pretraining dataset Omni-CT10K and the two downstream benchmarks (ExtBenchA and ExtBenchB) are collected from three independent centers. ExtBenchA is for in-domain training and evaluation. We assess 4 tasks, including *classification, segmentation, pathological TNM staging, and radiology report generation* on the test set. Implementation details for different tasks and data split are provided in the App. Sec. B. ExtBenchB is for assessing cross-center generalization, which consists of 495 patients (199 normal and 296 abnormal), among whom 292 have pathology reports. We use this dataset to evaluate *classification and TNM staging*. MSD: Finally, we evaluate the transferring ability of the representations on the public MSD *segmentation* dataset [28] which contains other types of organs and tumors, following the same training and test splits as in [29]. PLCC [30] and TriALS [31] assess cross-phase lesion segmentation capabilities. Details are provided in the App. Sec. B.

Evaluation metrics and baselines. Evaluation metrics: We report the AUC for classification and the Dice Similarity Coefficient (DSC) for segmentation. For radiology report generation, we use ROUGE-L and CIDEr, and a clinical efficacy metric, GREEN [32]. Baselines: To ensure a comprehensive comparison, we include representative methods spanning diverse pretraining paradigms, including self-supervised pretraining: VoCo [29] and SparK [26]; supervised pretraining: SuPreM [18] and UniModel [17]; vision-language pretraining: CLIP [33], fVLM [6], and Merlin [7]; and multi-modal supervised pretraining methods: VoComni [29], Triad [34], Argus [35], and Ark [36].

Implementation details. In Omni-CT10K, contrast-enhanced CT scans are registered with their corresponding non-contrast CT scans using the tools from [37, 38]. All volumetric data are preprocessed following the nnUNet protocol [39]. Our backbone is based on nnUNet [39] and consists of six encoder layers. For pretraining details, please refer to App. Sec. B.

Table 1: **Comparison with existing pre-training methods on ExtBenchA and MSD colon.**† denotes utilizing official weights; all others are re-implemented on Omni-CT10K and pre-training budget for fair comparison. The abbreviations Cls. and Seg. refer to Classification (Normal *vs.* Cancer) and Segmentation, respectively. * denotes copied from [29].

Method (Backbone)	Cls. (AUC %)	Seg. (DSC %)		Pathology (AUC %)				Report Generation (%)					MSD
		Organ	Tumor	T	N	M	Avg	BLEU-4	ROUGE-L	CIDEr	SPICE	GREEN	Colon
Train from scratch													
Swin [40]	90.55	92.90	45.75	62.23	58.05	39.25	53.18	0.01	21.62	2.21	18.58	32.34	35.13*
nnUNet [39]	85.44	94.41	53.82	55.94	51.40	30.89	46.08	0.61	23.62	2.19	21.64	20.59	41.69*
Self-supervised pretraining													
VoCov2 [29] (Swin)†	97.51	93.18	48.43	60.63	57.23	38.70	52.19	0.01	20.28	1.85	7.59	19.20	42.57*
SparK [25, 26, 11] (nnUNet)	95.80	94.51	53.95	57.30	52.53	41.99	50.61	0.01	26.83	1.83	28.97	14.48	44.47
Supervised pretraining													
SuPreM [18] (Swin)†	98.43	92.78	46.23	61.77	59.25	48.70	56.57	0.66	25.66	2.49	23.29	18.12	38.78*
SuPreM [18] (nnUNet)	97.39	94.85	58.30	63.91	57.26	53.42	58.20	0.01	27.29	1.97	24.43	24.34	48.42*
UniModel [17] (UNet)†	94.07	91.53	39.16	55.93	39.73	59.20	51.62	1.02	25.91	2.73	25.15	35.27	45.43
Vision-Language pretraining													
CLIP [41, 33] (nnUNet)	88.08	94.32	51.82	62.40	54.23	39.11	51.91	3.38	29.03	8.76	28.84	28.34	37.93
fVLM [6] (nnUNet)	97.28	94.18	54.47	63.74	52.05	11.44	42.41	1.97	28.41	4.78	27.41	22.31	41.76
Merlin [7] (ResNet)†	94.76	94.40	57.89	52.36	24.73	58.48	45.19	0.01	28.33	1.71	28.03	17.37	43.21
Multi-modal supervised pretraining													
VoComni [29] (nnUNet)†	98.50	93.87	48.97	59.84	53.15	41.37	51.45	0.01	28.47	1.90	27.59	15.06	43.85
Triad [34] (nnUNet)	98.06	94.54	54.44	64.63	58.17	35.62	52.81	3.30	28.54	11.15	29.58	32.70	45.61
Argus [35] (nnUNet)	98.14	94.85	56.24	66.24	56.52	48.15	56.97	5.36	32.65	16.47	31.52	34.57	49.43
Ark [36] (nnUNet)	98.23	94.98	61.62	66.46	60.43	**74.59**	67.16	0.13	19.17	0.43	21.81	22.05	50.22
Omni-supervised pretraining													
RadOmni (nnUNet)	**98.54**	**95.01**	**63.67**	**67.44**	**63.35**	73.56	**68.12**	**5.97**	**34.34**	**18.10**	**32.18**	**37.02**	**51.02**

4.2 Experiments on downstream tasks

In-domain performance across four downstream tasks. As shown in Tab. 1, existing pretraining paradigms often exhibit trade-offs across different objectives, whereas our approach achieves consistently strong performance across all four tasks, attaining new SOTA results. Notably, while multi-modal supervised pretraining methods perform well on classification, segmentation, and TNM staging, their performance on report generation remains limited. In contrast, RadOmni substantially improves report generation performance while maintaining leading results on the other tasks.

External-center zero-shot generalization. To evaluate zero-shot generalization, we further test the pre-trained model on ExtBenchB. The results are summarized in Tab. 2. RadOmni achieves the best AUC on tumor classification and pathology-related staging tasks, reaching 98.53 and 64.27, respectively. Compared with other pretraining approaches, our method maintains the best performance in both tasks, suggesting improved cross-center transferability under domain shift.

Transferring to other organs and tumor types. Despite being pretrained on Omni-CT10K, our method generalizes well to other unseen organs and tumor segmentation targets in the MSD benchmark. As shown in Tab. 3, it achieves the best performance in previously unseen Liver, Vessel, Pancreas, and Lung. The advantage is particularly pronounced on challenging targets such as Pancreas, where our

Table 2: **Zero-shot generalization on ExtBenchB.** † denotes utilizing official weights. RadOmni achieves the best generalization performance across both classification and pathology staging tasks.

Method	Cls. (AUC %)	Pathology (AUC %)			
		T	N	M	Avg
Train from scratch					
Swin [40]	78.97	69.39	54.03	50.39	57.94
nnUNet [39]	61.09	64.27	48.74	47.90	53.64
Self-supervised learning					
VoCov2 [29] (Swin)†	92.03	70.51	53.79	49.96	58.09
SparK [25, 26, 11] (nnUNet)	69.33	64.64	54.20	44.68	54.51
Supervised pretraining					
SuPreM [18] (Swin)†	90.51	66.75	58.33	48.62	57.90
SuPreM [18] (nnUNet)	81.80	67.22	64.04	47.89	59.72
UniModel [17] (UNet)†	68.37	66.32	54.51	51.49	57.44
Vision-language pretraining					
CLIP [41, 33] (nnUNet)	62.08	67.73	56.89	45.22	56.61
fVLM [6] (nnUNet)	88.12	64.67	50.47	44.76	53.30
Merlin [7] (ResNet)†	74.06	69.60	48.96	51.67	56.74
Multi-modal supervised pretraining					
VoComni [29] (nnUNet)†	93.97	67.37	51.94	44.25	54.52
Triad [34] (nnUNet)	92.01	72.97	58.94	43.87	58.59
Argus [35] (nnUNet)	96.81	71.00	60.32	**55.21**	62.18
Ark [36] (nnUNet)	97.58	70.51	63.45	52.01	61.99
Omni-supervised pretraining					
RadOmni (nnUNet)	**98.53**	**74.64**	**64.61**	53.54	**64.27**

Table 3: **Transfer to unseen MSD segmentation targets.** Most results are drawn from [29]. RadOmni shows the best transferability.

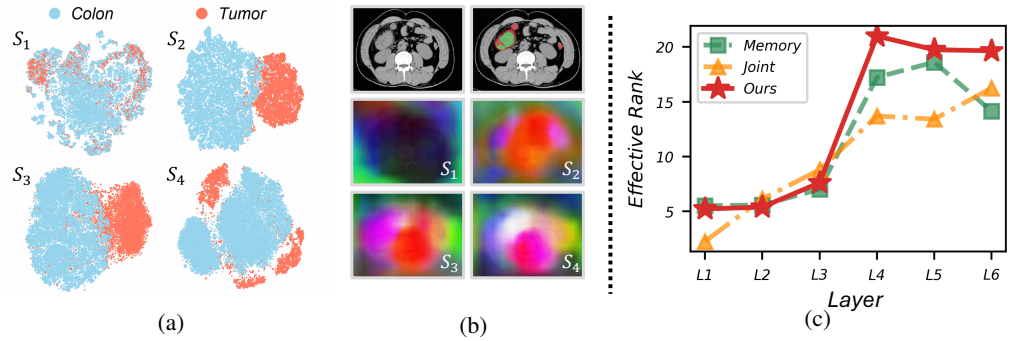
Method	MSD (%)				
	Liver	Vessel	Pancreas	Lung	Avg
Swin	67.85	64.80	47.88	70.28	62.70
nnUNet	67.43	68.92	46.08	64.73	61.79
SuPreM (Swin)	68.20	64.71	50.02	72.55	63.87
VoCo (Swin)	71.22	67.25	51.34	**75.74**	66.39
VoCo (nnUNet)	70.09	68.99	50.43	68.99	64.63
RadOmni (nnUNet)	**71.36**	**70.92**	**64.48**	74.38	**70.29**

Table 4: **Ablation results of curriculum stages.** For single-stage, no joint optimization with other stages are involved. As shown from Line 2-5, each stage provides complementary information. Aggregating all supervision resources achieves the best performance (Line 8). Results are reported on the validation set.

	Stage				Cls.	Tumor Seg.	Path. (T)
	S_1	S_2	S_3	S_4			
1					85.54	50.30	55.15
2	✓				95.75	54.07	56.37
3		✓			98.01	61.65	60.55
4			✓		96.07	51.08	55.06
5				✓	89.22	55.52	59.06
6	✓	✓			98.34	61.81	60.72
7	✓	✓	✓		98.38	61.68	61.12
8	✓	✓	✓	✓	**98.45**	**63.83**	**61.20**

Table 5: **Ablation results of different information-incorporation strategies.** Compared with providing supervision jointly or sequentially, or EMA strategy, which show limited performance gain in segmentation task, our method (with staged weight frozen) achieves the best performance across all tasks. Results are reported on the validation set.

Omni-method	Cls.	Tumor Seg.	Path. (T)
Baseline	85.54	50.30	55.15
Joint	97.55	56.76	60.38
Sequential	96.72	57.08	59.82
Memory (EMA)	98.00	61.54	60.55
Ours (w/o frozen)	98.41	55.59	60.76
Ours (w/ frozen)	**98.45**	**63.83**	**61.20**



(a) (b) (c)

Figure 3: **Qualitative ablation analysis across different pretraining stages and layers.** (a) t-SNE of NCCT representations at different stages, showing increasing separation between normal colon and tumor features, as well as among abnormal cases. (b) PCA-based feature visualization, where later stages exhibit clearer semantic and structural distinctions. (c) Effective-rank analysis across network layers. Compared with memory-based and joint-training strategies, our method maintains higher effective ranks in deeper layers, indicating improved representation diversity and capacity.

method surpasses VoCo by 14.05% and the vanilla nnUNet baseline by 18.40%. These results suggest that our omni-pretraining strategy learns transferable anomaly-aware localization across anatomically distinct organs, rather than overfitting to colon-specific patterns.

4.3 Ablation study

Ablation results of different curriculum stages. Quantitatively, table 4 shows that each individual stage provides useful supervision, while their contributions are clearly heterogeneous: S_1 (Line 2) offers a strong visual recognition, benefiting classification and segmentation, S_3 (Line 4) mainly enhances high-level semantic classification task, and S_4 (Line 5) directly enhances with the pathological task. More importantly, simply introducing more objectives does not reliably translate into better performance across tasks (Line 2-5). In contrast, progressively accumulating stages yields stronger transfer than isolated training, as shown in Line 6-8, validating that these sources of radiological knowledge are complementary when integrated in a proper training framework.

Qualitatively, the t-SNE visualizations in Fig. 3a shows a clear progression in representation quality of each stage. At the S_1 stage, normal organ and tumor features are not clearly separated and do not form stable clusters. S_2 introduces a strong coarse-grained separation, explicitly dividing normal organs and tumors into two clusters. However, the representation of different anomaly types are entangled together as a whole red tumor cluster. This is partially alleviated at the S_3 stage. The final S_4 further promotes this trend: normal organs stay clustered, whereas abnormal tumor features

remain differentiated under pathological information guidance. The PCA visualizations further verify the step-by-step identification of distinctive features (Fig. 3b).

Ablation results of different information-incorporation strategies. Quantitatively, the ablations in Tab. 5 (and Tab. 10 in App. Sec. C) explore the effectiveness of heterogeneous information aggregation strategies. Both joint and sequential training lead to clear degradation in segmentation performance, despite showing benefits on other tasks. The Memory-based method (EMA) partially alleviates this issue, but its gains remain inconsistent and marginal. In comparison, our method achieves best performance across different tasks, effectively preserving earlier visual knowledge while progressively injecting new objectives (also discussed in App. Sec. C).

Qualitatively, we further compare the internal quality of the learned representations by effective rank analysis, where higher rank means more meaningful representation [42]. As shown in Fig.3c, our method maintains consistently higher effective rank in deeper layers compared with memory-based and joint-training strategies, indicating improved representation diversity and capacity. In contrast, joint training exhibits substantially lower effective rank in deeper layers, suggesting optimization interference across heterogeneous objectives. We also provide CKA [43] analysis in App. Sec. C.

4.4 Data scaling analysis

As a foundation model, we present the pretraining data scaling behavior of RadOmni in Fig. 4. Compared with SSL and VLP, RadOmni consistently achieves superior AUC across all data scales and exhibits substantially stronger data efficiency. Notably, Omni already attains high performance in the low-data regime, indicating that heterogeneous clinical supervision provides highly informative representation priors even with limited pretraining data. Additionally, as the pretraining scale increases toward ∼10K samples, VLP begins to exhibit a steeper scaling trend, suggesting that vision-language pretraining relies more heavily on large-scale data to fully realize its representation capacity. Despite this accelerated improvement, our analysis indicates that VLP would require approximately ∼70K pretraining samples to achieve performance comparable to Omni trained with only 10K samples, highlighting the substantially higher data efficiency of Omni.

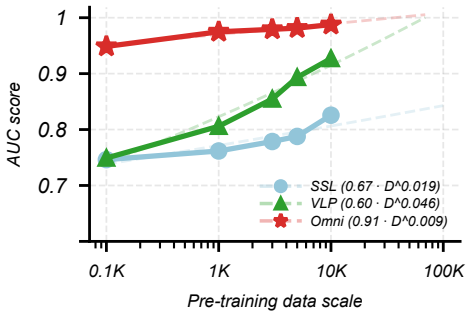


Figure 4: **Pretraining data scaling law.** RadOmni consistently outperforms SSL and VLP across all data scales. Although VLP shows a sharper scaling trend at larger data scales, it would require 7× data amount to achieve performance comparable to RadOmni, highlighting the scalability and efficiency of our design.

5 Conclusion and Discussion

In this work, we present RadOmni, a radiology omni-pretraining framework for non-contrast CT that progressively integrates heterogeneous clinical supervision. Built upon the Omni-CT10K dataset, which is composed by contrast-enhanced CT, segmentation masks, radiology reports, and pathology reports, RadOmni formulates omni-pretraining as progressive knowledge accumulation, where complementary clinical knowledge is introduced stage-by-stage through curriculum consolidation learning to mitigate optimization conflict and knowledge forgetting. Extensive experiments across internal, external, and public benchmarks demonstrate that RadOmni achieves strong and consistent performance across classification, segmentation, pathological staging, and report generation tasks, while exhibiting favorable scaling behavior, robust cross-center generalization, and transferable anomaly-aware representations. These results suggest that progressively integrating heterogeneous clinical knowledge is an effective strategy for improving the robustness, semantic capacity, and transferability of NCCT representations. More broadly, we hope this work encourages future research on unified medical pretraining frameworks leveraging heterogeneous clinical data to overcome the information limitations of medical imaging.

Limitations. Omni-CT10K provides a realistic testbed for radiology omni-pretraining, but its colorectal cancer focus limits claims of broad cross-organ generalization. Larger, more diverse

omni-modal datasets would enable more comprehensive evaluation. Computational constraints also prevented us from exploring larger vision models, which remains a direction for future work.

6 Ethics Statement

The study was approved by the ethics committee of Shanghai Institution of Pancreatic Disease (approval no. CHEC2026-189) and the ethics committees of all participating external centers. The requirement for informed consent was waived because of the retrospective study design. All data were de-identified before model training and testing.

7 Data, Code and Weights Availability

The source code, pretrained RadOmni weights, and de-identified Omni-CT10K imaging and associated clinical data are available upon reasonable request to the corresponding author, Jianpeng Zhang. Access to these resources is restricted to non-commercial academic research and requires institutional approval, IRB approval where applicable, and a signed agreement governing their use. Requests will receive a response within six weeks.

Acknowledgments and Disclosure of Funding

This work was supported by the Zhejiang Province Postdoctoral Research Excellence Funding Program (no. ZJ2024032), National Natural Science Foundation of China under Grant 62271465, National Key R&D Program of China under Grant 2025YFC3408300, Basic Research Program of Jiangsu (No. BK20255001), Suzhou Basic Research Program under Grant SYG202338, and the Sanming Project of Medicine in Shenzhen under Grant SZSM202411033. Jianpeng Zhang, Weiwei Cao, and Ling Zhang are employees of Alibaba DAMO Academy. Jianpeng Zhang and Ling Zhang hold Alibaba stock as part of the standard compensation package. The other authors declare no competing interests.

References

[1] David J Brenner and Eric J Hall. Computed tomography—an increasing source of radiation exposure. *New England journal of medicine*, 357(22):2277–2284, 2007.

[2] Daniel Thomas Ginat and Rajiv Gupta. Advances in computed tomography imaging technology. *Annual review of biomedical engineering*, 16(1):431–453, 2014.

[3] Wenya Linda Bi, Ahmed Hosny, Matthew B Schabath, Maryellen L Giger, Nicolai J Birkbak, Alireza Mehrtash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F Dunn, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: a cancer journal for clinicians*, 69(2):127–157, 2019.

[4] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[5] Kyongtae T Bae. Intravenous contrast medium administration and scan timing at ct: considerations and approaches. *Radiology*, 256(1):32–61, 2010.

[6] Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. In *The Thirteenth International Conference on Learning Representations*.

[7] Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, pages 1–11, 2026.

[8] Chiharu Sako, Brenda F Kurland, Taly G Schmidt, Jing He, Marius De Groot, Arpan A Patel, Dwight H Owen, Arya Amini, Brendan D Curti, Ronan J Kelly, et al. Deep-learning serial ct prediction of survival in immunotherapy-treated non–small cell lung cancer. *JAMA Network Open*, 9(1):e2555759, 2026.

[9] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[10] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 2577–2586, 2018.

[11] Fenghe Tang, Ronghao Xu, Qingsong Yao, Xueming Fu, Quan Quan, Heqin Zhu, Zaiyi Liu, and S Kevin Zhou. Hyspark: Hybrid sparse masking for large scale medical image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 330–340. Springer, 2024.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[13] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image classification and segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–6. IEEE, 2023.

[14] Fenghe Tang, Qingsong Yao, Wenxin Ma, Chenxu Wu, Zihang Jiang, and S Kevin Zhou. Hi-end-mae: Hierarchical encoder-driven masked autoencoders are stronger vision learners for medical image segmentation. *Medical Image Analysis*, page 103770, 2025.

[15] Fenghe Tang, Bingkun Nian, Yingtai Li, Zihang Jiang, Jie Yang, Wei Liu, and S Kevin Zhou. Mambamim: Pre-training mamba with state space token interpolation and its application to medical image segmentation. *Medical Image Analysis*, 103:103606, 2025.

[16] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023.

[17] 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.

[18] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised 3d models transfer to medical imaging tasks? In *The Twelfth International Conference on Learning Representations*.

[19] Jiacheng Hao, Xiaoming Zhang, Wei Liu, Xiaoli Yin, Yuan Gao, Chunli Li, Ling Zhang, Le Lu, Yu Shi, Xu Han, et al. Plus: Plug-and-play enhanced liver lesion diagnosis model on non-contrast ct scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 461–471. Springer, 2025.

[20] Zhiyun Zhang, Yan-Jie Zhou, Yujian Hu, Xiyao Ma, Zhouhang Yuan, Zirui Wang, Hongkun Zhang, and Minfeng Xu. Disentangling for transfer: Boosting limited modalities via information-theoretic regularization and cross-modal reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13052–13060, 2026.

[21] Chenhui Zhao, Yiwei Lyu, Asadur Chowdury, Edward Harake, Akhil Kondepudi, Akshay Rao, Xinhai Hou, Honglak Lee, and Todd Hollon. Towards scalable language-image pre-training for 3d medical imaging. *arXiv preprint arXiv:2505.21862*, 2025.

[22] Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23041–23050, 2025.

[23] Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony CW Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11238–11247, 2024.

[24] Rongsheng Wang, Fenghe Tang, Qingsong Yao, Rui Yan, Xu Zhang, Zhen Huang, Haoran Lai, Zhiyang He, Xiaodong Tao, Zihang Jiang, et al. Simcrop: Radiograph representation learning with similarity-driven cross-granularity pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 563–573. Springer, 2025.

[25] Keyu Tian, Yi Jiang, Chen Lin, Liwei Wang, Zehuan Yuan, et al. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. In *The Eleventh International Conference on Learning Representations*.

[26] Tassilo Wald, Constantin Ulrich, Stanislav Lukyanenko, Andrei Goncharov, Alberto Paderno, Maximilian Miller, Leander Maerkisch, Paul Jaeger, and Klaus Maier-Hein. Revisiting mae pre-training for 3d medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5186–5196, 2025.

[27] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[28] 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.

[29] Linshan Wu, Jiaxin Zhuang, and Hao Chen. Large-scale 3d medical image pre-training with geometric context priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[30] Jiawei Luo, Xiaoyu Wan, Jinchao Du, Li Liu, Ling Zhao, Xin Peng, Min Wu, Shixin Huang, and Xixi Nie. Comprehensive multi-phase 3d contrast-enhanced ct imaging for primary liver cancer. *Scientific Data*, 12(1):768, 2025.

[31] Marawan Elbatel, Mohamed Ghonim, Jiaji Mao, Zhuosheng Lin, Katharina Eckstein, Andrés Martínez Mora, Jonathan Deissler, Maximilian Rokuss, Constantin Ulrich, Zdravko Marinov, et al. Trials: Triphasic-aided liver lesion segmentation benchmark in non-contrast ct. *arXiv preprint arXiv:2605.16572*, 2026.

[32] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 374–390, 2024.

[33] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

[34] Shansong Wang, Mojtaba Safari, Qiang Li, Chih-Wei Chang, Richard LJ Qiu, Justin Roper, David S Yu, and Xiaofeng Yang. Triad: Vision foundation model for 3d magnetic resonance imaging. *arXiv preprint arXiv:2502.14064*, 2025.

[35] Weiming Zhuang, Chen Chen, Zhizhong Li, Sina Sajadmanesh, Jingtao Li, Jiabo Huang, Vikash Sehwag, Vivek Sharma, Hirotaka Shinozaki, Felan Carlo Garcia, et al. Argus: A compact and versatile foundation model for vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4418–4429, 2025.

[36] DongAo Ma, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. A fully open ai foundation model applied to chest radiography. *Nature*, 643(8071):488–498, 2025.

[37] Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–221. Springer, 2020.

[38] Tony CW Mok and Albert CS Chung. Conditional deformable image registration with convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2021.

[39] 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.

[40] 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.

[41] 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.

[42] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[43] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

[44] Chuyan Zhang, Hao Zheng, and Yun Gu. Dive into the details of self-supervised learning for medical image analysis. *Medical Image Analysis*, 89:102879, 2023.

[45] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172, 2020.

[46] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

[47] Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting rubik's cube: Self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 238–248. Springer, 2020.

[48] Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22873–22882, 2024.

[49] Xuefeng Ni, Linshan Wu, Jiaxin Zhuang, Qiong Wang, Mingxiang Wu, Varut Vardhanabhuti, Lihai Zhang, Hanyu Gao, and Hao Chen. Mg-3d: Multi-grained knowledge-enhanced vision-language pre-training for 3d medical image analysis. *Medical Image Analysis*, page 104027, 2026.

[50] Jianghui Wang, Yang Chen, Xingyu Xie, Cong Fang, and Zhouchen Lin. Task-robust pre-training for worst-case downstream adaptation. *Advances in Neural Information Processing Systems*, 36:9458–9478, 2023.

[51] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[52] Ruobing Huang, J Alison Noble, and Ana IL Namburete. Omni-supervised learning: scaling up to large unlabelled medical datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 572–580. Springer, 2018.

[53] Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. Multi-modal vision pre-training for medical image analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5164–5174, 2025.

[54] Xingxin He, Aurora Rofena, Ruimin Feng, Haozhe Liao, Zhaoye Zhou, Albert Jang, and Fang Liu. Omnimri: A unified vision–language foundation model for generalist mri interpretation. *arXiv preprint arXiv:2508.17524*, 2025.

[55] Nahid Ul Islam, DongAo Ma, Jiaxuan Pang, Shivasakthi Senthil Velan, Michael Gotway, and Jianming Liang. Foundation x: integrating classification, localization, and segmentation through lock-release pretraining strategy for chest x-ray analysis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3647–3656. IEEE, 2025.

[56] Meilin Liu, Jiaying Wang, and Jing Shan. Omnifm: Toward modality-robust and task-agnostic federated learning for heterogeneous medical imaging. *arXiv preprint arXiv:2603.21660*, 2026.

[57] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[58] Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12):3033–3043, 2023.

[59] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 2002.

[60] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

[61] Colin Studholme, Derek LG Hill, and David J Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.

[62] 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):e230024, 2023.

[63] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Appendix.

A Related work.

Representation learning in medical image analysis. The acquisition of high-quality representations is fundamental to the success of diverse medical imaging applications [4]. Recently, *self-supervised learning* (*SSL*) has emerged as a important paradigm for label-efficient representation learning [44, 45]. Existing methodologies primarily rely on contrastive learning [46, 47, 48, 29, 40] and masked image modeling [14, 11, 12, 13, 15]; however, these approaches often encounter significant challenges in achieving superior scaling and generalization capabilities [44]. To facilitate effective scaling, researchers have explored diverse supervision paradigms. Among these, *vision-language pretraining* (*VLP*), which leverages the reports to enhance visual representations, is a promising representation learner [7, 6, 21, 24, 49, 23]. Nevertheless, the semantics learned through these methods often fail to support fine-grained representation learning. Although recent frameworks such as fVLM [6] and SimCroP [24] have attempted to address this limitation, the performance of these models in dense prediction tasks remains suboptimal. Furthermore, while *supervised pretraining* (*SP*) offers a scalable alternative for representation learning [17, 18, 16], it is inherently constrained by the nature of the supervision. Consequently, the utility of these models is often restricted to downstream tasks that align closely with the upstream supervision, preventing the achievement of robust performance across a comprehensive range of general tasks [50]. To overcome the aforementioned limitations and achieve general, scalable representations, we propose RadOmni which integrates the strengths of diverse data modalities and representation learning paradigms and its omni representations that achieve superior performance across a wide spectrum of tasks

Medical omni-supervised pretraining. Recent medical foundation models have highlighted the value of large-scale pretraining under heterogeneous or multi-modal supervision [34, 51]. Existing efforts can be broadly grouped into three lines. The first focuses on large-scale representation learning from heterogeneous datasets or mixed supervision, including VoComni [29], which combines labeled, unlabeled, and pseudo-labeled CT data, Ark [36], which aggregates heterogeneous labels across datasets, and OSL [52], which scales a small labeled set to a much larger unlabeled medical dataset through iterative pseudo-label refinement. The second line studies multimodal representation learning, including Merlin [7], which learns CT representations from paired CT, radiology reports, and diagnosis codes, BrainMVP [53], which models cross-modal correspondence in multi-modal MRI, and OmniMRI [54], which unifies image-only, vision-text, and instruction-response supervision in MRI. The third line emphasizes task or system unification, including Foundation X [55], which integrates classification, localization, and segmentation through heterogeneous task annotations, and OmniFM [56], which targets modality- and task-agnostic federated learning under heterogeneous imaging environments. However, these pipelines do not jointly model clinically co-occurring supervision from non-contrast CT, enhanced CT, radiologist annotations, radiology reports, and pathology reports in a unified pretraining framework. To address this gap, we curate Omni-CT10K, a 10K-scale CT dataset from unique patients, and propose RadOmni, which performs omni-pretraining for non-contrast CT through progressive knowledge injection from these complementary modalities.

B Implementation Details

Details of Omni-CT10K. For Omni-CT10K, all tumor cases are confirmed by pathology and/or surgery, whereas normal cases are verified by colonoscopy. All CT scans were acquired preoperatively, and the interval between CT acquisition and pathology report did not exceed three months. Among these, 3,930 abnormal cases and 2,040 normal cases have corresponding contrast-enhanced CT. Within the contrast-enhanced abnormal subset, 3,443 cases are further accompanied by pathology reports. Table 7 summarizes patient characteristics and CT acquisition parameters on Omni-CT10K. Paired NCCT and CECT scans come from the same patient and the same examination. CECT availability depends on clinical indications. Cross-phase supervision uses valid paired scans, with registration quality assessed in Tab. 11. We use Qwen3-Max [27] to extract organ-related descriptions and TNM-related descriptions from CT reports and pathology

Table 6: Details of nnUNet backbone.

Layer	Feature size	Kernel size	Stride
1	32	$[1,3,3]$	$[1,1,1]$
2	64	$[1,3,3]$	$[1,2,2]$
3	128	$[3,3,3]$	$[1,2,2]$
4	256	$[3,3,3]$	$[2,2,2]$
5	320	$[3,3,3]$	$[2,2,2]$
6	320	$[3,3,3]$	$[2,2,2]$

Table 7: Patient and acquisition distributions in Omni-CT10K.

Age (years)	< 50 1,820	50–64 4,314	65–74 3,101	≥ 75 1,356
Sex	Male 6,133	Female 4,452	Other 6	– –
Scanner manufacturer	Philips 4,718	Toshiba 3,598	Siemens 1,898	United Imaging 377
Tube voltage (kVp)	< 100 28	100–119 0	120 10,547	> 120 16
Tube current (mA)	< 200 41	200–299 1,357	300–399 6,327	≥ 400 2,866
CTDIvol (mGy)	< 10 53	10–14.9 3,053	15–19.9 7,469	≥ 20 16
Slice thickness (mm)	≤ 1 4,248	1.5–3 966	5 5,376	> 5 1

reports, respectively. The prompt templates are shown in Fig. 5 and Fig. 6. To assess extraction quality, a radiologist manually reviewed a randomly sampled subset of 2,106 reports ($\sim$20%). The review assessed two criteria: faithfulness, whether the extracted text introduced clinically relevant information absent from the source report; and coverage, whether all colorectal-related descriptions in the source report were retained. Of the reviewed reports, 2,105 (99.95%) satisfied faithfulness, 2,094 (99.43%) satisfied coverage, and 2,094 (99.43%) satisfied both criteria.

Details of ExtBenchA. It comprises 3,682 unique patients, including 2,291 normal cases and 1,391 abnormal cases, with manually annotated organ and tumor masks as well as 1,391 pathology reports. We randomly split the dataset into training/finetuning, validation, and test sets with a ratio of 6:2:2.

Pretraining implementation details. For pretraining, $\underline{S_1}$: we follow the same training setting as [11], with all encoder layers trainable. $\underline{S_2}$: the encoder is initialized with the pretrained weights from S_1, and all encoder layers remain trainable. $\underline{S_3}$: we freeze the first two layers of the visual encoder to preserve shallow generic representations. $\underline{S_4}$: we freeze the first four layers of the visual encoder to preserve shallow and intermediate representations. In S_3, and S_4, a pretrained BERT [57] is used as the text encoder. Stages S_1, S_3, and S_4 are trained for 100 epochs on 8 NVIDIA H20 GPUs with a total batch size of 48. Stage S_2 is trained for 1000 epochs on a single NVIDIA H20 GPU with a batch size of 2. Tab. 9 compares the GPU hours and peak memory of RadOmni and joint training.

Table 8: Overview details of data processing and omni pre-training settings.

Data pre-processing (nnUNet)	
Spacing	$5.0 \times 0.736 \times 0.736$ (mm)
Non-contrast CT	Intensity = $[-1001, 88]$ std = 298.98 mean = -145.13
Contrast-enhanced CT	Intensity = $[-3024, 121]$ std = 375.55 mean = -148.73
Omni pre-training settings (S_1)	
Epoch	100
Optimizer	AdamW
Optimizer LR	1e-4
Weight decay	1e-5
Batch size	48
LR schedule	cosine decay
Sub-volume size	$32 \times 192 \times 256$
Omni pre-training settings (S_2)	
nnUNet default setting	-
Sub-volume size	$32 \times 192 \times 256$
Omni pre-training settings (S_3, S_4)	
Epoch	100
Optimizer	AdamW
Optimizer LR	1e-4
Weight decay	5e-2
Global batch size	$8 \times 6 = 48$
LR schedule	warmup cosine
Warmup steps	1000
Input-volume size	$48 \times 256 \times 384$

The details of nnUNet backbone are shown in Tab. 6. The details of our data processing settings and per-training are shown in Tab. 8. In S_2, the encoder is initialized from S_1, while the decoder is randomly initialized. For the contrast-enhanced CT, we train a teacher network with the same architecture for segmentation and then frozen. In S_3 and S_4, we apply linear projections to map the multi-scale visual descriptors to 256 dimension, and employ distinct attention layers to extract query representations for report and TNM prototype alignment, respectively. The number of attention heads and the dropout rate are set to 4 and 0.1, respectively. In

Table 9: Pretraining computational costs of joint training and RadOmni. GPU-hours are reported for each stage and summed across stages. Peak GPU memory is reported in GB; the overall peak is the maximum across stages.

Metric	Strategy	S1	S2	S3	S4	Total / Peak
GPU-hours	Joint training	882	–	–	106	988
	RadOmni	882	4	91	104	1081
Peak GPU memory (GB)	Joint training	58.6	–	–	78.8	78.8
	RadOmni	58.6	4.9	61.2	57.6	61.2

Example Prompt Template for Radiology Reports

[Role]
You are a professional radiologist specializing in CT imaging.

[Task]
Please read the following CT report and extract only the original descriptions related to the colorectum, then output them in the same or der as they appear in the report.

The colorectum scope includes, but is not limited to, the colon, rectum, sigmoid colon, transverse colon, ascending colon, descending colon, ileocecal region, appendix.

Colorectum-related descriptions include, but are not limited to, mass, space-occupying lesion, tumor, polyp, cauliflower-like lesion, thick ening, stenosis, obstruction, dilatation, gas accumulation, fluid accumulation, air-fluid level, perforation, free gas, ischemia, necrosis, inflammation, diverticulum or diverticulitis, appendicitis, blurred surrounding or peribowel fat planes, exudation, enlarged mesenteric or peribowel lymph nodes.

[Output Rules]
Output only the original text, without rewriting, summarizing, or inferring.
If multiple colorectum-related descriptions appear in the same paragraph, list them one by one.
If the full report contains no colorectum-related description, note that descriptions such as no abnormality in the colorectum. In that case, output only: No colorectum-related description.

[Output Format]
Output only the extracted original colorectum-related descriptions. Do not output any other explanatory text.

Figure 5: Example prompt template for extracting organ-related descriptions from CT reports.

addition, the contrastive loss in S_2 is optimized only on tumor-positive cases, while S_4 is performed only on tumor-positive cases with pathology reports.

Downstream implementation details. For classification task, following [58], we perform patch-level prediction. Specifically, a cropped patch is labeled as 1 if it contains a tumor and as 0 otherwise. During inference, we apply a sliding-window strategy and use the maximum score across all patches as the final prediction. For segmentation task, we use the nnUNet framework [39]. For the pathological TNM staging task, we crop sub-patches according to tumor location, and predict pathological TNM result using three separate classification heads. For the report generation task, all experiments integrate their corresponding frozen pre-trained image encoder with the same trainable Transformer-based text decoder. The generation process of our method and all comparison methods is optimized using the language modeling loss [57] for 100 epochs on 8 NVIDIA H20 GPUs. For experiments on pretraining data scaling, we report the average results on ExtBenchA and ExtBenchB.

Implementation details of baseline methods. For all baseline methods, we follow the official settings and retrain each method using the corresponding modalities from Omni-CT10K. Specifically, for SparK, we use only non-contrast CT. For SuPreM, we use organ segmentation as the supervised training objective. For Traid, and CLIP, we use only non-contrast CT and reports. For fVLM, we use NCCT, reports, and segmentation labels. For Argus and Ark, we retrain the models using all modalities, which incurs the same training cost as RadOmni.

Implementation details of S_2. For $\mathcal{L}_{cont}$, the adjacent contextual organ region is derived directly from the manual segmentation label. Let M_{tum} and M_{colon} denote the binary tumor and colon masks, respectively. We first construct a ring-shaped neighborhood around the tumor by taking the dilation of M_{tum}:

$$M_{ring} = \mathrm{clip}(\mathrm{Dilate}(M_{tum}; k) - M_{tum},\, 0,\, 1)\,,$$

and define the contextual mask as the intersection of this shell with the colon region,

$$M_{ctx} = \mathrm{clip}(M_{ring} \odot M_{colon},\, 0,\, 1)\,.$$

Example Prompt Template for Pathology Reports

[Role]
You are an expert in oncologic pathology report interpretation and clinical text extraction.

[Task]
From the input pathology report, extract all statements relevant to the TNM framework and group them into:

T-related descriptions, referring to the primary tumor and its local extent or invasion;
N-related descriptions, referring to regional lymph node evaluation or nodal metastasis;
M-related descriptions, referring to distant metastatic disease.

For T-related descriptions, include statements about tumor invasion, depth of infiltration, involved anatomical layers, direct extension int o adjacent tissues or organs, and other descriptions reflecting local tumor extent.
For N-related descriptions, include statements about regional lymph node examination, the presence or absence of nodal metastasis, i nvolved lymph nodes, and extranodal extension.
For M-related descriptions, include statements about distant metastasis, metastatic involvement of distant organs or non-regional sites, explicit absence of distant metastasis, or inability to assess metastatic disease.

The goal is to recover TNM-relevant textual evidence from the pathology report rather than to infer or assign a TNM stage.

[Output Rules]
Use only information explicitly stated in the report.
Do not infer missing findings or assign TNM stage labels.
Preserve both positive and negative findings.
Classify direct extension into adjacent structures as T-related.
Classify regional lymph node disease as N-related.
Classify distant metastatic disease as M-related.
Do not misclassify lymphovascular invasion or perineural invasion as nodal or distant metastasis.
Combine all relevant descriptions within each category into a single concise statement.
If a category is absent, output "Not mentioned."

[Output Format]

```
{
    "T-related descriptions": "...",
    "N-related descriptions": "...",
    "M-related descriptions": "..."
}
```

Figure 6: Example prompt template for extracting TNM descriptions from pathology reports.

By default, the dilation parameters are set to $k = 11$. For each feature level z_i, the masks are resized to match the spatial resolution of the corresponding feature map. The tumor and context prototypes are computed via masked average pooling. To avoid degenerate supervision, $\mathcal{L}_{cont}$ is evaluated only at feature levels where both the tumor mask and the corresponding contextual mask remain non-empty after resizing.

C Further experiment analysis

Comparison analysis. The comparison in Tab. 1 across pretraining paradigms highlights why existing methods struggle to achieve strong performance on all four benchmarks. Self-supervised methods provide only limited overall gains and remain weak on high-level semantic tasks such as pathological TNM staging and report generation. Supervised pretraining improves certain tasks, e.g., SuPreM reaches 58.30 DSC on tumor segmentation, but still underperforms on language-intensive tasks. Vision-language pretraining shows the opposite tendency, performing better on report generation but lagging behind on dense prediction, with MSD colon tumor DSC scores consistently below ours. Multimodal pretraining methods are more balanced, yet none achieves strong results across all tasks. Overall, the main challenge lies in mitigating forgetting and optimization conflicts across heterogeneous objectives while maintaining structural, semantic, and cross-task compatibility.

For generalization, as shown in Tab. 2, the comparison across pretraining paradigms reveals that generalization tends to improve with stronger supervision. SSL and SP can benefit specific tasks but often show uneven transfer, while VLP exhibits less stable generalization. In contrast, multi-modal and omni-pretraining methods deliver consistently stronger results, with Argus, Ark, and RadOmni all showing robust tumor classification generalization. Among them, our method achieves the best overall balance across classification and pathology prediction, indicating the most stable cross-task transfer to unseen centers.

Information incorporation strategies and information forgetting analysis. As shown in Tab. 10, our method preserves stable and largely monotonic improvements across stages, whereas sequential training, memory-based strategies, and the non-frozen variant exhibit noticeable degradation after

Table 10: Forgetting analysis of different omni-pretraining strategies across four tasks. Background green / yellow indicates increase/decrease from previous stage.

Method	Cls. (AUC %)			Seg. (DSC %)			Path. (AUC %)			R.G. (GREEN [32])		
	S_2	S_3	S_4	S_2	S_3	S_4	S_2	S_3	S_4	S_2	S_3	S_4
Sequential	98.36	98.18	96.72	60.99	58.86	57.08	59.90	57.71	59.87	36.31	41.06	18.83
Memory	98.02	98.14	98.23	59.42	64.27	61.62	63.79	61.11	67.44	24.62	22.32	22.05
Ours (w/o freeze)	98.36	98.35	98.41	60.99	53.63	56.49	59.90	66.80	64.61	36.31	24.57	36.15
Ours (w freeze)	98.36	98.50	98.54	60.99	65.11	63.67	59.90	67.18	68.12	36.31	30.07	37.02

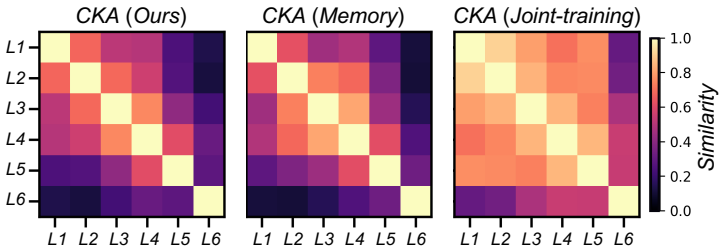


Figure 7: CKA [43] analysis of different information incorporation methods.

introducing later supervision. In particular, sequential training suffers substantial performance drops in report generation after pathology supervision is introduced (41.06 → 18.83), indicating severe forgetting of previously acquired semantic knowledge. Memory-based training partially alleviates this issue, but still shows inconsistent optimization across tasks. Similarly, removing layer freezing leads to degradation in segmentation performance after semantic supervision is introduced (60.99 → 53.63), suggesting interference with previously learned anatomical representations. In contrast, our full strategy consistently maintains or improves performance across classification, segmentation, pathology prediction, and report generation, demonstrating that selective freezing and progressive joint optimization are important for stabilizing heterogeneous knowledge incorporation and mitigating knowledge forgetting during omni-pretraining.

CKA analysis. Fig. 7 presents the inter-layer CKA similarity analysis of different information incorporation strategies. Our method exhibits progressively decreasing similarity between distant layers, indicating clearer functional differentiation across network depth. In contrast, joint training produces uniformly high inter-layer similarity, particularly among intermediate layers, suggesting representation homogenization caused by optimization interference across heterogeneous objectives. The memory-based strategy partially alleviates this issue, but still shows stronger coupling between adjacent layers compared with our method. These results suggest that curriculum consolidation encourages more hierarchical and specialized representations, while mitigating representational collapse during heterogeneous omni-pretraining.

Registration Analysis. We evaluate registration quality using MI [59, 60], NMI [61], and SSIM, and further assess its practical impact by applying TotalSegmentator [62] to predict the colon mask of CECT and NCCT. And then we compute the DSC score between the corresponding masks before and after registration contrast-enhanced scan against the non-contrast. As shown in Table 11, registration improves consistently, with NMI increasing from 1.15 to 1.24, MI from 0.87 to 1.29, SSIM from 0.43 to 0.63, and DSC from 71.41% to 89.81%. These results demonstrate that registration not only enhances image-level similarity but also leads to substantially better anatomical overlap.

Effective analysis of cross-phase transfer in S_2. We further validate the effectiveness of introducing the cross-phase transfer capability of CECT. As shown in Table 12, we conduct an ablation study by

Table 11: Evaluation performance of registration

	NMI	MI	SSIM	DSC (%)
w/o register	1.15	0.87	0.43	71.41
w/ register	1.24	1.29	0.63	89.81

Table 12: Ablation study of cross-phase transfer on MSD dataset.

	Colon	Liver	Vessel	Pancreas	Lung	Avg
w/o CECT	49.42	71.50	**70.18**	74.32	73.08	67.70
w/ CECT	**51.29**	**72.02**	70.06	**74.62**	**74.07**	**68.41**

Table 13: Ablation study of component in S_2.

Align	$\mathcal{L}_{KL}$	MSD Colon
		41.69
	✓	50.23
$\mathcal{L}_{cont}$		48.13
$\mathcal{L}_{mse}$	✓	47.99
$\mathcal{L}_{mmd}$ [63]	✓	50.35
$\mathcal{L}_{cont}$	✓	**51.29**

comparing an NCCT lesion and organ segmentation-only setting (i.e., w/o CECT) with a cross-phase transfer setting that incorporates CECT (i.e., w/ CECT). The results show that the introduction of contrast-enhanced CT improves not only the segmentation of the target colon tumor but also the segmentation of tumors at other anatomical sites. Across the five MSD segmentation tasks, the average DSC increases by 0.71. Additionally, the ablation results on MSD Colon (shown in Table 13) show that all introduced objectives improve upon the baseline of 41.69, although the magnitude of the gains varies across alignment strategies. $\mathcal{L}_{KL}$ alone increases the score to 50.23, and its combination with $\mathcal{L}_{cont}$ further yields the best result of 51.29, outperforming other KL-based combinations, including $\mathcal{L}_{mse}$ and $\mathcal{L}_{mmd}$. These results indicate that the combination of KL and contrastive alignment provides an effective and complementary design. In addition, We assess the transferability of RadOmni's pretrained representations across three complementary scenarios. First, on additional NCCT datasets, RadOmni improves lesion Dice from 56.02% to 59.06% on PLCC [30] and from 33.42% to 38.60% on TriALS [31] (Tab. 17 and Fig. 8).

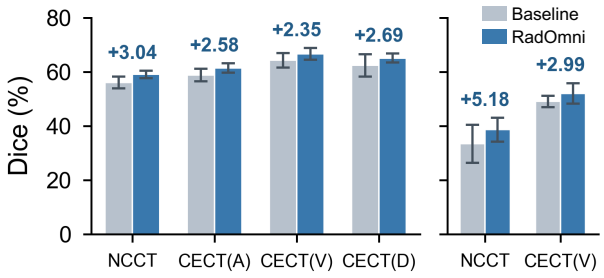


Figure 8: **Cross-phase capabilities of RadOmni on PLCC (Left) and TriALS (Right) lesion segmentation datasets.**

Stable anatomical segmentation across S_3 and S_4. As shown in Table 14, the segmentation-based anatomical parsing remains stable in both stages. Colon segmentation achieves 97.78% in S_3 and 97.68% in S_4, with standard deviations of 1.85 and 1.88, indicating consistent accuracy and variance across the two late stages.

Sensitivity analysis across specificity thresholds To further evaluate the classification performance under clinically relevant constraints, we investigate the sensitivity of the proposed method at high specificity requirements of 90%, 95%, and 99%, denoted as $SE@90SP$, $SE@95SP$, and $SE@99SP$, respectively. The experimental results, summarized in Table 15, demonstrate that our method achieves competitive diagnostic sensitivity on the internal test set, with $SE@90SP$ and $SE@95SP$ reaching 96.97% and 92.59%. At the most stringent $SE@99SP$ threshold, our model maintains a sensitivity of 75.08%, which is comparable to omni-supervised approaches VoComni and Ark. Notably, our framework exhibits favorable generalization on the external dataset, yielding 95.61% for $SE@90SP$ and 89.19% for $SE@99SP$, while prior methods such as Ark and VoComni

Table 14: Stability evaluation of anatomical structure parsing

Colon segmentation (%)			
S_3		S_4	
mean	std	mean	std
97.78	1.85	97.68	1.88

Table 15: Comparison with existing pre-training methods on classification tasks. Sensitivities at different specificity of 90%, 95%, and 99% are reported on both the internal and the external test sets,, denoted as $SE@90SP$, $SE@95SP$, and $SE@99SP$, respectively. † denotes utilizing official weights; all others are re-implemented under same Omni-CT10K setting for fair comparison.

Method (Backbone)	Internal (%)			External (%)		
	$SE@90SP$	$SE@95SP$	$SE@99SP$	$SE@90SP$	$SE@95SP$	$SE@99SP$
Train from scratch						
Swin [40]	94.28	84.85	61.95	72.97	61.49	40.88
nnUNet [39]	50.17	24.24	10.77	15.54	4.05	2.70
Self-supervised learning						
VoCov2 [29] (Swin)†	94.61	89.90	80.13	88.51	82.43	66.89
SparK [25, 26, 11] (nnUNet)	87.88	78.45	51.18	27.03	20.95	6.08
Supervised pretraining						
SuPreM [18] (Swin)†	96.30	91.58	72.39	75.34	63.85	54.39
SuPreM [18] (nnUNet)	91.58	84.85	56.23	61.82	45.27	27.70
UniModel [17] (UNet)†	80.81	68.01	40.74	22.30	13.18	3.72
Vision-Language pretraining						
CLIP [41, 33] (nnUNet)	60.61	45.12	9.76	18.92	10.81	2.36
fVLM [6] (nnUNet)	93.94	84.51	40.74	71.28	57.77	47.64
Merlin [7] (ResNet)†	87.21	74.75	35.35	39.53	27.70	11.15
Multi-modal supervised pretraining						
VoComni [29] (nnUNet)†	96.63	92.27	75.11	83.11	77.70	71.62
Triad [34] (nnUNet)	93.94	90.91	67.34	76.69	75.34	59.12
Argus [35] (nnUNet)	94.95	89.90	**76.43**	90.88	85.81	79.05
Ark [36] (nnUNet)	95.29	90.91	74.07	93.24	90.88	82.77
Omni-supervised pretraining						
RadOmni (nnUNet)	**96.97**	**92.59**	75.08	**95.61**	**92.57**	**89.19**

Table 16: Comparison with existing pre-training methods on colon-segment segmentation tasks. † denotes utilizing official weights; all others are re-implemented under same Omni-CT10K setting for fair comparison.

Method (Backbone)	Colon-Segment segmentation (DSC %)					
	Ascending	Transverse	Descending	Sigmoid	Rectum	Avg
Train from scratch						
Swin [40]	91.88	88.91	87.00	85.20	88.78	88.36
nnUNet [39]	94.38	92.29	90.87	88.64	91.11	91.46
Self-supervised learning						
VoCov2 [29] (Swin)†	92.37	89.63	88.34	86.08	89.30	89.14
SparK [25, 26, 11] (nnUNet)	94.40	92.35	91.08	88.65	91.22	91.54
Supervised pretraining						
SuPreM [18] (Swin)†	94.70	92.89	91.28	89.39	91.60	91.99
SuPreM [18] (nnUNet)	91.85	88.41	87.15	84.58	88.64	88.13
UniModel [17] (UNet)†	87.83	83.25	82.03	79.98	86.25	83.87
Vision-Language pretraining						
CLIP [41, 33] (nnUNet)	94.23	92.02	90.73	88.45	90.82	91.25
fVLM [6] (nnUNet)	93.87	91.50	90.33	88.12	90.69	90.90
Merlin [7] (ResNet)†	93.98	91.84	90.49	88.16	90.44	90.98
Multi-modal supervised pretraining						
VoComni [29] (nnUNet)†	93.15	90.77	89.22	86.98	89.00	89.82
Triad [34] (nnUNet)	94.31	92.31	90.96	88.46	91.16	91.44
Argus [35] (nnUNet)	94.64	92.46	90.81	89.11	91.43	91.69
Ark [36] (nnUNet)	94.73	92.93	**91.42**	89.52	91.72	92.06
Omni-supervised pretraining						
RadOmni (nnUNet)	**94.85**	**92.97**	91.30	**89.54**	**91.74**	**92.08**

show a more noticeable performance gap under these metrics. Compared to baseline configurations like the train-from-scratch nnUNet, which experiences a sharp decline to 15.54% $SE@90SP$ on external data, the proposed pretraining strategy effectively mitigates the impact of domain shift.

Table 17: Cross-phase segmentation on PLCC and TriALS. Dice is reported as mean ± standard deviation under five-fold cross-validation. Baseline denotes nnUNet without pretraining.

		Liver		Lesion	
Dataset	Phase	Baseline	RadOmni	Baseline	RadOmni
PLCC	NCCT	91.79 ± 1.55	$\mathbf{92.12} \pm 1.58$	56.02 ± 2.19	$\mathbf{59.06} \pm 1.38$
	Arterial	93.05 ± 1.33	$\mathbf{93.24} \pm 1.34$	58.82 ± 2.32	$\mathbf{61.40} \pm 1.70$
	Venous	93.84 ± 1.42	$\mathbf{93.87} \pm 1.45$	64.25 ± 2.70	$\mathbf{66.60} \pm 2.19$
	Delayed	93.91 ± 1.67	$\mathbf{94.06} \pm 1.63$	62.41 ± 4.15	$\mathbf{65.10} \pm 1.66$
TriALS	NCCT	–	–	33.42 ± 7.04	$\mathbf{38.60} \pm 4.42$
	Venous	–	–	49.07 ± 2.12	$\mathbf{52.06} \pm 3.76$

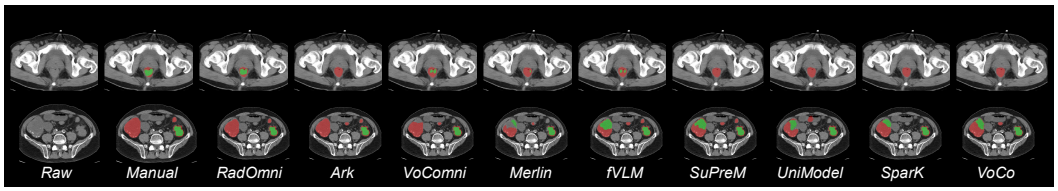


Figure 9: Quantitative evaluation of organ (red) and tumor (green) segmentation.

Evaluation segmentation on colon-segment. To verify the effectiveness of our approach, we conduct experiments on an internal dataset specifically focused on five colon segmentation segmentation (*i.e.* ascending colon, transverse colon, descending colon, sigmoid colon, and rectum), which annotated by professional radiologists, and we follow the same data split as benchmarks. The segmentation results are shown in Tab. 16, our RadOmni achieves an average DSC of 92.08%, outperforming various pretraining paradigms including self-supervised, supervised, and vision-language methods. Specifically, our approach yields superior results in four out of five anatomical regions, notably reaching 94.85% in the ascending colon and 89.54% in the sigmoid colon, the latter being a challenging area for most baselines. Compared to the baseline trained from scratch (91.46%), our method demonstrates robust capability in capturing complex morphological features across the intestinal tract.

D Visualization

We qualitatively evaluate the segmentation performance through the visualizations presented in Fig. 9 and Fig. 10. Specifically, we observe the capabilities of various pretraining paradigms and methods in the task of non-contrast-enhanced CT tumor segmentation. As illustrated in these examples, methods based on self-supervised learning (SSL) appear to struggle with generalization in certain complex cases, while methods based on vision-language pretraining (VLP) are observed to generate more false positives. In contrast, Omni-based approaches tend to suppress false positives more effectively. Within the Omni framework, the proposed method is shown to achieve relatively more precise lesion boundaries. Furthermore, we extend the evaluation to the MSD multi-organ tumor segmentation task (Fig. 10). The promising results obtained on this benchmark further validate that the proposed method learns highly transferable, anomaly-aware representations, thereby demonstrating its robustness across diverse scenarios.

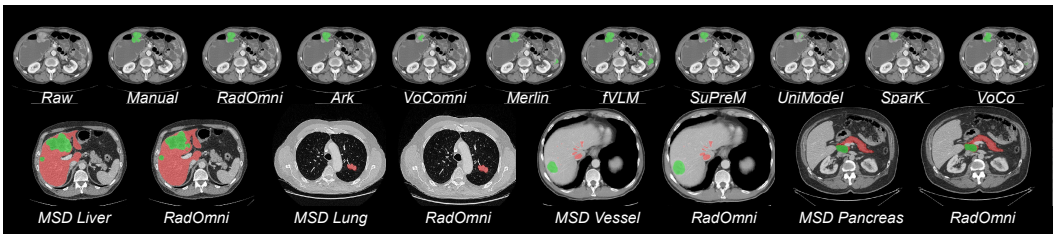


Figure 10: Quantitative evaluation of multiple organ and tumor segmentation on MSD benchmark.

E Declaration of LLM Usage

We used Qwen3-Max to extract organ-related descriptions from radiology reports and TNM-related descriptions from pathology reports. These extracted descriptions provide textual supervision for stages S3 and S4, respectively. The extraction procedures and prompt templates are described in Appendix B and Fig 5 and 6. Clinicians reviewed a randomly selected subset of the extracted descriptions to assess their quality. LLMs were also used to improve the clarity and readability of the manuscript.

F Societal Impacts

Positive societal impacts. RadOmni may positively impact society by improving the diagnostic utility of NCCT, a widely accessible, low-cost, and safer imaging modality in routine clinical practice. By transferring complementary knowledge from contrast-enhanced CT, radiology reports, and pathology reports into NCCT representation learning, the proposed framework may support more accurate and generalizable computer-aided diagnosis, especially in settings where contrast imaging or expert interpretation is limited. This could help reduce diagnostic disparities across institutions and improve clinical efficiency in tasks such as classification, segmentation, staging, and report generation.

Negative societal impacts. Potential negative societal impacts include the risk of incorrect or biased predictions if the model is deployed without sufficient clinical validation across diverse populations, institutions, and imaging protocols. Errors in diagnosis, staging, or report generation could adversely affect clinical decision-making and patient safety. In addition, the use of large-scale multimodal medical data raises concerns regarding privacy, data security, and governance. There is also a risk that users may over-rely on automated outputs in ways that exceed the validated scope of the model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims reflect the paper's contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in the Conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in Appendix B and additional experiment result in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We provide detailed experimental settings and implementation descriptions in Appendix B. The data, source code, and pretrained model weights are available upon reasonable request to the corresponding author, as described in Section 7.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe the experimental setup in the main paper and provide full implementation and training details in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our method outperforms baseline approaches by a substantial margin in most settings. Additionally, we provide comprehensive experiments and ablation studies. Section C reports mean Dice scores and standard deviations across five cross-validation folds on PLCC and TriALS.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B specifies the detailed description of compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We adhere to the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential societal impacts in Appendix F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited the data and models used in the paper, and we have properly credited the original creators while complying with the corresponding licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Appendix B documents the Omni-CT10K dataset and the model architecture and training settings. Documentation accompanying the data, source code, and pretrained weights is provided through the controlled-access process described in Section 7.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This study uses retrospectively collected, de-identified clinical data and does not involve crowdsourcing, direct participant recruitment, or participant-facing experimental procedures. Participant instructions and compensation are therefore not applicable. Ethical approval and the waiver of informed consent are described in Section 6.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study was approved by the ethical committees of Shanghai Institution of Pancreatic Disease (approval No. CHEC2026-189) and all the external centers participated in this study. Participant consent was waived due to the retrospective nature of this study. All data were de-identified prior to model training and testing.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use Qwen3-Max to extract organ-related and TNM-related descriptions from clinical reports for supervision in stages S3 and S4. This use is disclosed in Appendix E, with implementation details and prompt templates provided in Appendix B and Figures 5 and 6.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.