Versatile Vision-Language Model for 3D Computed Tomography

Jiayu Lei[1,2], **Ziqing Fan**[2,3], **Yanyong Zhang**[1], **Weidi Xie**[2,3], **Ya Zhang**[2,3,4*], **Yanfeng Wang**[2,3*]

[1]University of Science and Technology of China, Anhui, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3]School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China
[4] Institute of Artificial Intelligence for Medicine, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

Abstract

Representation learning serves as a foundational component of medical vision-language models (MVLMs), enabling cross-modal alignment, semantic consistency, and enhanced generalization capabilities for downstream tasks. As generalist models rapidly evolve, there is a pressing need to unify diverse downstream tasks, such as diagnosis, segmentation, report generation, and multiple choice within a cohesive framework, demanding more efficient and versatile visual representation learning. However, current MVLMs predominately follow CLIP-style vision pretraining, failing to leverage heterogeneous data resources with multi-dimensional imaging and diverse annotation forms. And there lacks systematic analysis of efficient vision encoder design across varied downstream applications, including diagnosis, segmentation, and text generation tasks, particularly for volumetric imaging like Computed Tomography (CT). Besides, current MVLMs exhibit constrained voxel-level capabilities, lacking effective multi-task instruction tuning framework capable of achieving robust performance across various downstream tasks. To address these challenges, we propose **CTInstruct**, a novel MVLM employing a hybrid ResNet-ViT encoder with multi-granular vision-language pretraining for efficient heterogeneous data modeling, and unified instruction tuning that jointly optimizes discriminative, generative, and voxel-level reasoning for volumetric medical imaging. CTInstruct achieves SOTA performance across 8 CT benchmarks, setting a new standard for data-efficient multimodal learning in medical imaging.

Code — https://github.com/ljy19970415/CTInstruct

1 introduction

Medical imaging is a cornerstone of modern healthcare, supporting accurate disease diagnosis, lesion characterization, and treatment planning. Recent advancements at the intersection of artificial intelligence (AI) and radiology have been accelerated by the convergence of computer vision and large language models (LLMs)(Bai et al. 2025; Alayrac et al. 2022; Liu et al. 2023; Hurst et al. 2024; Ramesh et al. 2021; Fan et al. 2025a; Lee et al. 2021), laying the foundation for the development of medical vision-language models (MVLMs)(Wu et al. 2023; Li et al. 2023; Lei et al. 2024; Moor et al. 2023; Tu et al. 2024; Fan et al. 2025b).

*Corresponding Authors.

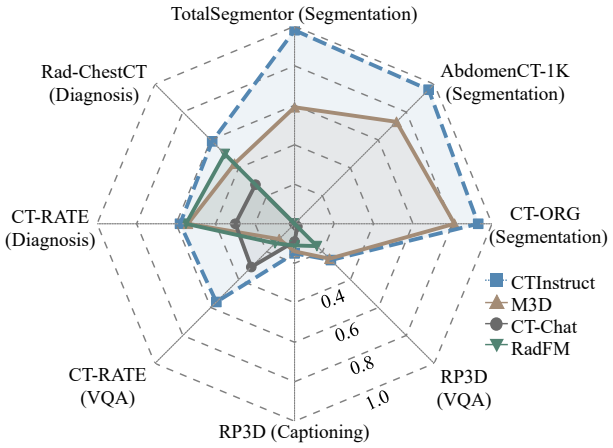

Figure 1: CTInstruct sets SOTA on 8 public benchmarks for segmentation (DSC), diagnosis (AUC), and generation (METEOR) tasks versus current 3D generative MVLMs.

Despite rapid progress, existing medical vision-language models—such as LLaVA-Med (Li et al. 2023), Med-Flamingo (Moor et al. 2023), and RadFM (Wu et al. 2023)—are primarily built upon vision-language contrastive learning frameworks and are optimized for global semantic tasks like question answering, captioning, or report generation. However, these models face several limitations. First, their vision encoder architectures and pretraining strategies are not designed to effectively leverage heterogeneous data sources, including both multi-dimensional (2D and 3D) imaging and diverse supervision signals such as diagnostic labels, segmentation masks, and free-text reports. This limitation is particularly problematic in the context of medical imaging, where annotated data is scarce and task diversity is high. Second, their downstream performance is constrained—especially on spatially grounded tasks such as segmentation—due to coarse-grained vision-language pretraining and the absence of robust multi-task instruction tuning frameworks. As a result, existing MVLMs struggle to generalize across tasks that require fine-grained spatial understanding and task-specific reasoning.

In this paper, we introduce **CTInstruct**, a vision-language model for 3D CT that addresses the limitations of exist-

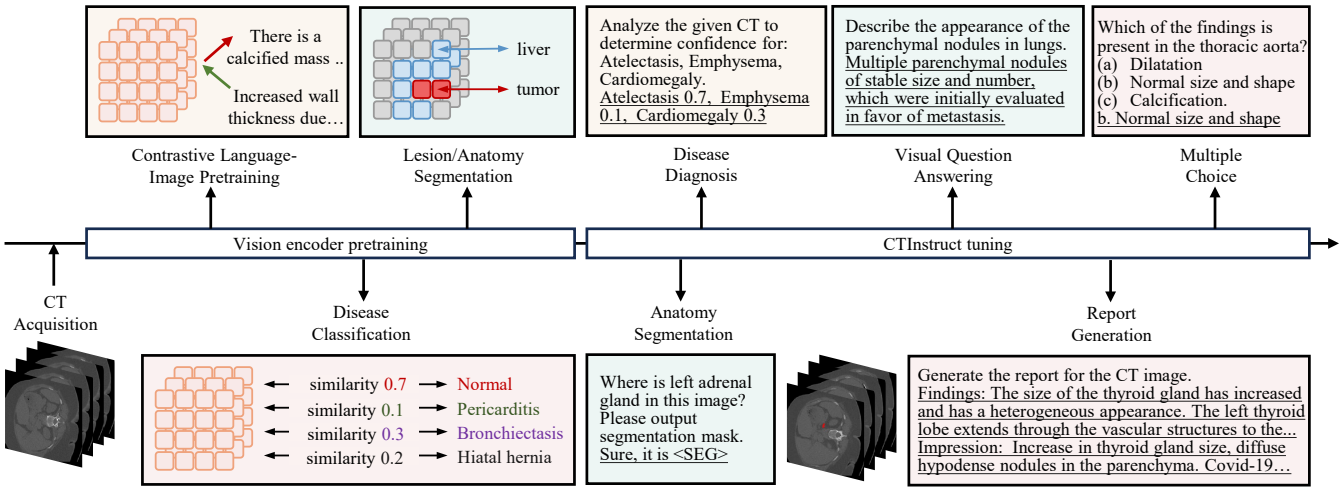

Figure 2: The proposed CTInstruct framework integrates multi-granular visual representation learning with multi-task instruction tuning for 3D CT analysis. Our vision encoder captures rich vision-language alignment patterns through pretraining on heterogeneous data, including both multi-dimensional imaging (2D and 3D) and diverse supervision signals. During instruction tuning, CTInstruct enables unified multi-task optimization across classification, segmentation, and generation tasks.

ing MVLMs. It employs a hybrid ResNet-ViT encoder with multi-granular pretraining to leverage heterogeneous data, including 2D/3D images, labels, masks, and reports. A unified instruction-tuning framework with task-specific tokens enables joint optimization across classification, segmentation, and generation tasks. CTInstruct effectively captures both semantic and spatial information, achieving SOTA performance on eight public benchmarks as shown in Figure 1.

Specifically, for vision encoder pre-training, we leverage manually annotated medical imaging datasets curated by the community (Hamamci et al. 2024a; Wu et al. 2023; Draelos et al. 2021a; Lei et al. 2025), and unify a diverse set of tasks within a coherent image-text alignment framework. In contrast to existing MVLMs, which predominantly rely on global contrastive learning over large-scale image-text pairs, our approach explicitly integrates spatial and semantic objectives into the vision encoder training. Specifically, the pretraining incorporates three complementary objectives: volume-level alignment with the dense clinical description, disease classification aligned with category-level textual descriptions, and voxel-wise segmentation for fine-grained spatial understanding. This joint optimizations allow the model to simultaneously capture high-level semantic and fine-grained spatial context, which is critical for volumetric medical data. By leveraging shared textual embeddings across tasks, our alignment mechanism facilitates knowledge transfer between tasks with varying granularity, thereby improving generalization and ensuring compatibility with heterogeneous label spaces. Following pretraining, we introduce a unified instruction-following paradigm that enables CTInstruct to jointly perform diagnostic reasoning, semantic segmentation, and generative tasks. Through the use of task-specific tokens, the model enables to interpret and respond to diverse prompts within a single framework.

In summary, we make the following contribution:

1. **Multi-Granular Vision Pretraining**: We propose the first vision encoder pretraining strategy that jointly optimizes volume-level semantic alignment, disease classification, and voxel-wise segmentation using heterogeneous data resources, enabling holistic spatial-semantic feature learning for volumetric medical data.
2. **Extensive Vision Encoder Ablations for MVLM**: Our extensive ablations reveal that a hybrid ResNet-ViT trained with multi-granular alignment on heterogeneous data maximizes downstream performance, achieving data-efficient medical representation learning.
3. **CTInstruct**: A novel MVLM that unifies discriminative and generative capabilities through multi-granular pretraining leveraging heterogeneous CT annotations and multi-task instruction tuning. In the experiment, CTInstruct achieves new SOTA results across 8 CT benchmarks spanning segmentation, diagnosis, and generation.

2 Related Works

Table 1 comprehensively compares generative MVLMs and specialized models for CT modality across key dimensions: input data dimensionality, vision encoder architecture, pretraining strategy, and downstream applications. Early generative MVLMs like LLaVA-Med (Li et al. 2023) and Med-Flamingo (Moor et al. 2023) utilize 2D ViT encoders trained via image-text contrastive learning for medical visual question answering (VQA). However, their inability to process 3D volumes limits applicability to CT data. RadFM (Wu et al. 2023) further unifies 2D/3D training through depth-wise replication of 2D inputs processed with 3D ViT. Subsequent approaches transitioned to 3D-native architectures: Both CT-Chat (Hamamci et al. 2024b) and M3D (Bai et al. 2024) employ CLIP-pretrained 3D ViT encoders. While CT-Chat focuses exclusively on CT-RATE-derived VQA tasks,

Methods		Input Dimension		Vision Encoder		Pretraining Strategy			Downstream Application		
		2D	3D	ViT	ResNet	ITC	Seg.	Clas.	Diagnosis	VQA	Segmentation
Generative MVLMs	LLaVa-Med	✓		✓		✓			✓	✓	
	Med-Flamingo	✓		✓		✓			✓	✓	
	RadFM	✓	✓	✓		✓			✓	✓	
	CT-Chat		✓	✓		✓			✓	✓	
	M3D		✓	✓		✓			✓	✓	✓
	CTInstruct (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Specialized Models	CT-CLIP		✓	✓		✓			✓		
	CT-Net		✓		✓			✓	✓		
	MedSAM	✓		✓			✓				✓
	SegVol		✓	✓			✓				✓

Table 1: Overview of the recent works on MVLMs. ✓ denotes the application of each method regarding input dimension, vision encoder architecture, pretraining strategy, and supported downstream applications. (Seg.: Segmentation, Clas.: Classification)

M3D integrates text generation and segmentation in instruction tuning but shows limited segmentation capability, likely due to its coarse-grained vision pretraining. Beyond generative MVLMs, researches in specialized models target specific clinical tasks. For diagnosis, CT-Net (Draelos et al. 2021b) adopts cross-entropy-pretrained ResNet, while CT-CLIP (Hamamci et al. 2024a) advances this with CLIP-pretrained ViT for text-guided classification. In segmentation, prompting-driven frameworks like MedSAM (Ma et al. 2024a) and SegVol (Du et al. 2024) enable segmentation across hundreds of anatomical categories.

3 Methodology

We design CTInstruct as the optimal configuration revealed by experimental results. Below, we introduce our problem statement, model design, and two-stage training strategy.

3.1 Problem Statement

As stated in Figure 2, our goal is to obtain a comprehensive medical vision-language model for 3D Computed Tomography, which can handle most medical needs including both low level tasks (lesion/anatomy segmentation) and high level tasks (diagnosis, visual question answering, and report generation). Mathematically, given a dataset $\mathcal{D} = \{\mathcal{X}_n, \mathcal{L}_n\}_{n=1}^{N}$ where $\mathcal{X}_n$ represents a CT image and $\mathcal{L}_n = \{l_n^j | l_n^j \in \{S, C, T\}\}$ denotes a set of multimodal labels (S: segmentation labels, C: classification labels, T: textual descriptions), the objective is to train a generative MVLM through a visual question-answering paradigm:

$$y = \Phi(\mathcal{X}, \mathcal{Q}), \tag{1}$$

where $\Phi(\cdot)$ denotes the model, $\mathcal{Q}$ represents a task-specific query (*e.g.*, requesting a diagnostic result, clinical report, or anatomical region localization), and y denotes the corresponding output (*e.g.*, disease confidence score, clinical report, or segmentation mask).

3.2 Vision Encoder Pretraining

Capitalizing on the remarkable success of Large Language Models (LLMs), a critical focus in developing effective generative MVLMs lies in the vision encoder design. Its architecture and training strategies are essential for aligning complex visual features with linguistic embeddings, especially for high-dimensional 3D volumes like CT. This section introduces our vision encoder framework, designed to leverage heterogeneous clinical data through two key innovations: (i) A hybrid ResNet-ViT architecture unifying 2D/3D representation learning by jointly modeling inter-slice contextual relationships and intra-slice spatial features, and (ii) Multi-granular image-text alignment pretraining facilitating knowledge transfer across image-text pairing, diagnosis, and segmentation tasks via shared textual embeddings. [1]

Vision Encoder Architecture. The input 2D/3D medical imaging first undergoes any-resolution preprocessing (Xue et al. 2024) to preserve native resolution while optimizing memory consumption, then decomposes into volumetric patches $X \in \mathbb{R}^{P \times h \times w \times d}$, where (h, w, d) represents the spatial dimensions of each patch and P denotes the total patch count. Our vision encoder employs a hybrid architecture for input image processing, comprising three integrated components as shown in Figure 3: (i) a ResNet2D slice encoder extracting intra-slice spatial features within individual patch, (ii) a patch-wise attention module modeling contextual relationships among patches, and (iii) a slice-wise attention module capturing inter-slice dependencies within individual patch. Specifically, the input X are reshaped into stacked 2D slices $\mathbb{R}^{Pd \times h \times w}$ and processed through the ResNet2D encoder, yielding multi-resolution feature maps $\phi_{\text{vision}}(X)$. The final layer (m_{th} layer) feature map is restruc-

[1] Training data: CT-RATE (Hamamci et al. 2024a), RP3D (Wu et al. 2023), INSPECT (Huang et al. 2023), Rad-ChestCT (Draelos et al. 2021a), CC-CCII (Nguyen et al. 2021), UCSD-AI4H (Zhao et al. 2020), SARS-COV (Soares et al. 2020), RadImageNet (Mei et al. 0), COVID-19 (Jun et al. 2020), Abdomen1K (Ma et al. 2021), CT-ORG (Rister et al. 2020), TotalSegmentor (Wasserthal et al. 2023). ATM22 (Zhang et al. 2023), BTCV (Landman et al. 2015), AMOS22 (Ji et al. 2022), MM-WHS-CT (Zhuang and Shen 2016), LUNA16 (Setio et al. 2017), FLARE22 (Ma et al. 2024b), WORD (Luo et al. 2022), RibFrac (Yang et al. 2025), MSD (Antonelli et al. 2022), NSCLC (Bakr et al. 2018), KiTS23 (Heller et al. 2023), AutoPET (Gatidis et al. 2022), SegThor (Lambert et al. 2020), SEGA (Radl et al. 2022), PARSE22 (Luo et al. 2023), Pancreas CT (Roth et al. 2015), LNDB (Pedrosa et al. 2021), KIPA22 (He et al. 2021), Hector22 (Andrearczyk et al. 2022), FUMPE (Masoudi et al. 2018), DAP (Jaus et al. 2023), CT-Pelvic (Liu et al. 2021).

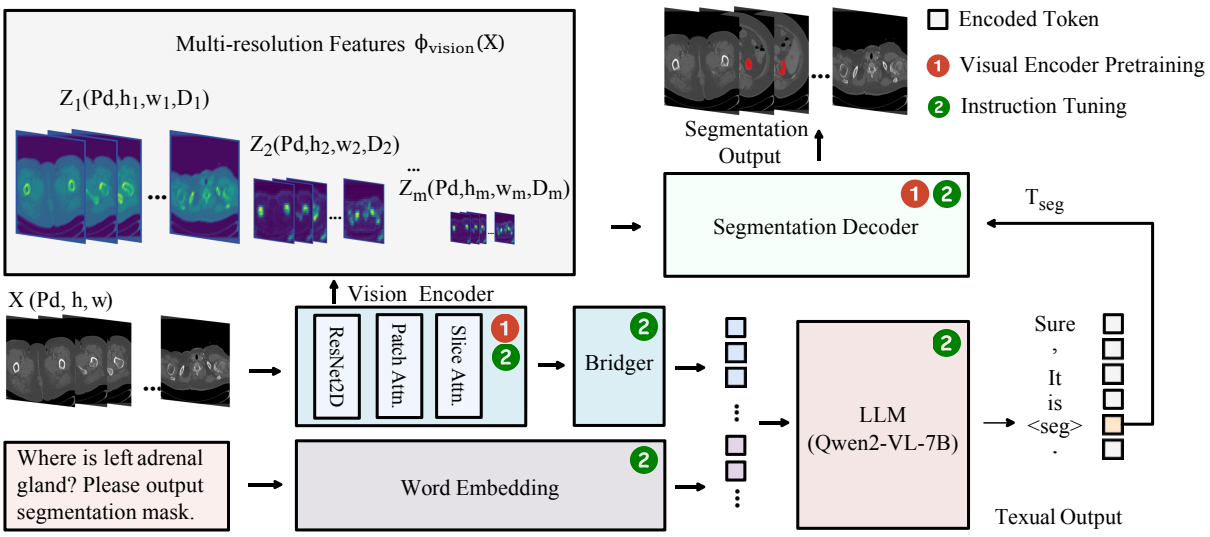

Figure 3: Overview of our framework. The system consists of vision encoder, bridger module, segmentation decoder, and LLM.

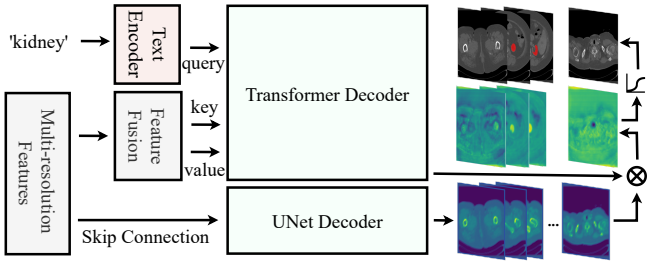

Figure 4: Segmentation decoder architecture.

tured into $\mathbb{R}^{d\times Ph_mw_m\times D_m}$ for patch-wise attention where D_m is the feature dimensionality of m_{th} layer. Subsequent slice-wise attention operates on $\mathbb{R}^{Ph_mw_m\times d\times D_m}$ to model inter-slice correlations within patches. The resulting post-attention feature $\mathbb{R}^{Pd\times h_m\times w_m\times D_m}$ undergoes spatial average pooling to form the image representation $Z\in\mathbb{R}^{Pd\times D_m}$:

$$Z=\frac{1}{hw}\sum_{i}^{h}\sum_{j}^{w}(\phi_{\text{vision}}(X)_m)_{i,j}, \tag{2}$$

where (i,j) denotes spatial position within the feature map.

Segmentation Decoder Architecture. To enable segmentation in both vision encoder pretraining and instruction tuning, we incorporate a segmentation decoder $\phi_{\text{decoder}}(\cdot)$ consisting a U-Net architecture for spatial feature map generation, and a transformer decoder for text-prompted segmentation implementations (Zhao et al. 2023). Specifically: (i) The U-Net decoder's final layer produces feature maps $\mathbb{R}^{Pd\times h\times w\times D_1}$ via skip connections from multi-resolution feature maps $\phi_{\text{vision}}(X)$, where D_1 denotes the last first feature dimensionality; (ii) category names are encoded into text embeddings $T_{\text{seg}}\in\mathbb{R}^{C\times D_{\text{text}}}$ serving as transformer decoder queries, where C is category amount and D_{text} is the textual embedding dimensionality; (iii) multi-resolution feature maps $\phi_{\text{vision}}(X)$ undergo fusion and projection:

$$Z_f=\text{project}(\text{concat}(\text{pool}(\phi_{\text{vision}}(X)))), \tag{3}$$

yielding $Z_f\in\mathbb{R}^{h_mw_m\times Pd\times D_{\text{text}}}$ as keys and values of transformer decoder; (iv) Finally, the voxel-text alignment is accomplished via dot product between the projected transformer decoder outputs $\mathbb{R}^{Pd\times C\times D_1}$ and the last layer U-Net feature map $\mathbb{R}^{Pd\times h\times w\times D_1}$, yielding output masks $\mathbb{R}^{P\times C\times h\times w\times d}$.

Vision Encoder Pretraining Task Design. At the first training stage, we unify a wide range of vision-language alignment tasks under a cohesive multitask learning framework. Unlike prior works (Wu et al. 2023; Bai et al. 2024; Hamamci et al. 2024b,a) that primarily focus on image-text contrastive learning, we uniquely incorporate vision-language alignment patterns from classification and segmentation datasets. The core pretraining tasks are as follows:

- **Image-text contrastive pretraining.** The classic image-text contrastive pretraining paradigm is adopted (Remy, Demuynck, and Demeester 2022):

$$\mathcal{L}_{\text{CLIP}}=-\frac{1}{N}\sum_{i=1}^{N}\left(\ln\frac{e^{Z_i\cdot T_i/\tau}}{\sum_{j=1}^{N}e^{Z_i\cdot T_j/\tau}}+\ln\frac{e^{T_i\cdot Z_i/\tau}}{\sum_{j=1}^{N}e^{T_i\cdot Z_j/\tau}}\right),$$

  where Z is the post-attention image feature and T is the text embedding. This pretraining establishes clinical text-image correlations, enhancing cross-modal representations for generative tasks.
- **Category description alignment pretraining.** Inspired by vision encoder pretraining techniques of specialized diagnostic models (Zheng et al. 2024; Draelos et al. 2021b), we adopt a category alignment task:

$$p_i^c=\frac{Z\cdot T_{\text{cls}}^c}{\|Z\|\times\|T_{\text{cls}}^c\|},$$
$$\mathcal{L}_{\text{cls}}=L_{\text{bce}}(p_i,\hat{p}_i). \tag{4}$$

where Z is post-attention visual feature and T_{cls} is text embedding via modality-specific projection heads. This pretraining strategy aligns visual patterns with clinical narratives by maximizing the cosine similarity between input image and text embedding of the correct category.

- **Voxel-text alignment pretraining.** We also integrate voxel-level classification capabilities essential for fine-grained medical perception from segmentation tasks by optimizing through dice and binary cross entropy loss:

$$y_i = \phi_{\text{decoder}}(Z, T_{\text{seg}}, \phi_{\text{vision}}(X)),$$
$$\mathcal{L}_{\text{seg}} = L_{\text{bce}}(y_i, \hat{y}_i) + L_{\text{dice}}(y_i, \hat{y}_i).$$

It boosts the spatial reasoning of pretrained vision encoders through voxel-level anatomical and pathological vision-language alignment, enhancing performance on downstream generative tasks requiring spatial awareness.

Finally, the vision encoder simultaneously optimizes multi-granular alignment objectives:

$$\mathcal{L}_{\text{mtp}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{CLIP}}. \tag{5}$$

3.3 Vision-language Instruction Tuning

In this section, we introduce the Bridger and LLM architecture, then present our unified approach for multi-task tuning using task-specific instructions.

Bridger and LLM Architecture. We employ a lightweight linear projection layer as the vision-language bridger (Li et al. 2023), mapping image embeddings $Z \in \mathbb{R}^{Pd \times D_m}$ into the textual latent space $Z' \in \mathbb{R}^{Pd \times D_{\text{text}}}$. Unlike fixed-length tokenization schemes implemented in Perceiver architectures (Li et al. 2022), our approach preserves a dynamic token count determined by the volumetric patch decomposition parameters (P and d), which correspond to the original input shapes. This design maximizes retention of fine-grained visual information while maintaining discrimination capacity across varied imaging inputs. For LLM, we leverage the language part of Qwen2-VL-7B-Instruct (Bai et al. 2025) for its proficiency in linguistic understanding and instruction-following.

Instruction Tuning Task Design. Our instruction tuning incorporates a comprehensive clinical task spectrum including open generation tasks, segmentation decoding, and diagnostic confidence prediction within a unified instructional paradigm. We first process raw sample and desired answer into Qwen2VL's instruction template as follows:

> **Instruction of open generation:**
> <Image>...<Image> Please generate a report.
> **Desired answer:** Findings: Minimal pleural effusion ...
> **Instruction of semantic segmentation:**
> <Image>...<Image> Where is the COVID infection?
> **Desired answer:** Sure, it is <SEG>.
> **Instruction of diagnosis:**
> <Image>...<Image> Give probabilistic predictions for the presence of calcification and bronchiectasis.
> **Desired answer:** calcification 0.9, bronchiectasis 0.2

The detailed training losses are as follows:

- **Open generation tasks.** The loss for the generated tokens follow the classic next token prediction loss (Radford et al. 2018):

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log P(y_t | \mathbf{y}_{<t}, \mathbf{x}), \tag{6}$$

where x denotes the multimodal input, y_t is the target token at position t, and T is the sequence length.

- **Segmentation decoding.** For semantic segmentation tasks, $\mathcal{L}_{\text{LM}}$ trains the model to predict <SEG> token. The final-layer embedding of this token serves as T_{seg} input of $\phi_{\text{decoder}}(\cdot)$, which generates segmentation masks optimized through $\mathcal{L}_{\text{seg}}$ (Lai et al. 2024).
- **Diagnostic confidence prediction.** For diagnosis tasks, $\mathcal{L}_{\text{LM}}$ trains the model to predict a <disease_confidence> token. The final-layer embedding of this token feeds into a classification head that outputs disease confidence scores, optimized via $\mathcal{L}_{\text{cls}}$.

Finally, we jointly optimize the vision encoder, bridger, segmentation decoder, and LLM using the multi-task instruction dataset spanning report generation, captioning, segmentation, diagnosis, multiple-choice, and finding-related QA.

4 Experiments

We begin this section by detailing the experimental setup. Next, we present ablation studies identifying the optimal vision encoder design. Finally, we benchmark CTInstruct against state-of-the-art MVLMs.

Task Type	Dataset	Training	Testing
Diagnosis	CT-RATE	24128	1564
	Rad-ChestCT	3188	360
Vision Question-Answering	RP3D-Cap	88891	1991
	RP3D-VQA	464341	14001
	CT-RATE-VQA	1198107	38145
Segmentation	CT-ORG	546	134
	Abdomen1K	3180	770
	TotalSegmentor	14144	1585
	COVID19	17	3

Table 2: Dataset Information.

4.1 Experimental Setup

Datasets. We conduct experiments on 9 public datasets: CT-RATE (Hamamci et al. 2024a), Rad-ChestCT (Draelos et al. 2021a), RP3D-Cap, RP3D-VQA (Wu et al. 2023), CT-RATE-VQA (Hamamci et al. 2024a), COVID-19 (Jun et al. 2020), TotalSegmentor (Wasserthal et al. 2023), Abdomen1K (Ma et al. 2021), and CT-ORG (Rister et al. 2020). All datasets adhere to their official train-test splits. Table 2 shows task types and the number of training and testing data.

Metrics. The image-text retrieval task is evaluated using IR@5, IR@10, TR@5, TR@10 across 200 samples. The diagnosis task is evaluated using standard multi-label classification metrics: the Area Under the ROC Curve (AUC),

Pretrain Strategy			RP3D				CT-RATE				Rad-ChestCT				COVID-19	TotalSegmentor	CT-ORG	Abdomen1K
ITC	Seg.	Clas.	IR@5	IR@10	TR@5	TR@10	IR@5	IR@10	TR@5	TR@10	AUC	ACC	F1	Pre.	DSC	DSC	DSC	DSC
✓			3.50	6.00	3.50	7.00	3.50	6.50	4.50	7.00	54.73	55.88	44.10	35.27	68.83	98.04	85.56	91.36
✓		✓	3.00	7.00	4.00	8.50	6.50	9.50	3.50	8.00	55.07	56.73	43.78	39.02	87.08	98.32	90.78	93.14
✓	✓		6.00	9.00	4.50	7.50	5.00	9.00	5.50	9.00	56.82	57.79	43.95	33.20	91.56	98.44	93.30	93.69
✓	✓	✓	4.50	8.00	5.00	7.50	5.00	10.00	4.50	9.00	57.44	62.89	44.75	35.62	92.33	98.58	93.38	93.72

Table 3: Vision encoder pretraining strategy ablation with 3D pretraining data and Hybrid ResNet-ViT encoder. Darker grid shading indicates better outcomes. All values are percentages (%). (Seg.: Segmentation, Clas. Classification)

Vision Encoder			RP3D				CT-RATE				Rad-ChestCT				COVID-19	TotalSegmentor	CT-ORG	Abdomen1K
Res.	ViT.	Hyb.	IR@5	IR@10	TR@5	TR@10	IR@5	IR@10	TR@5	TR@10	AUC	ACC	F1	Pre.	DSC	DSC	DSC	DSC
✓			4.00	6.00	2.00	7.00	5.00	10.50	4.50	9.00	54.76	57.60	43.67	33.03	77.46	98.00	83.84	89.10
	✓		1.50	6.50	2.00	5.50	6.50	14.00	6.50	12.50	56.73	57.51	43.26	33.80	80.18	98.41	88.17	93.16
		✓	4.50	8.00	5.00	7.50	5.00	10.00	4.50	9.00	57.44	62.89	44.75	35.62	92.33	98.58	93.38	93.72

Table 4: Vision encoder architecture ablation with 3D pretraining data and multitask pretraining. Darker grid shading indicates better outcomes. All values are percentages (%). (Res.: ResNet3D, ViT.: ViT3D, Hyb.: Hybrid ResNet-ViT)

Data Dimension		RP3D				CT-RATE				Rad-ChestCT				COVID-19	TotalSegmentor	CT-ORG	Abdomen1K
3D	2D+3D	IR@5	IR@10	TR@5	TR@10	IR@5	IR@10	TR@5	TR@10	AUC	ACC	F1	Pre.	DSC	DSC	DSC	DSC
✓		4.50	8.00	5.00	7.50	5.00	10.00	4.50	9.00	57.44	62.89	44.75	35.62	92.33	98.58	93.38	93.72
	✓	5.00	9.00	4.50	9.00	4.50	11.00	5.50	10.00	57.22	61.33	46.09	38.46	93.35	99.00	94.74	95.07

Table 5: Data Dimensionality ablation with multitask pretraining and Hybrid ResNet-ViT encoder. Darker grid shading indicates better outcomes. All values are percentages (%).

Accuracy (ACC), Precision (Pre.), and F1-Score (F1). Segmentation is quantified via the Dice Similarity Coefficient (DSC). For open generative tasks, we adopt BLEU-1, ROUGE-1, and METEOR.

4.2 Vision Encoder Pretraining Strategy Ablation

To systematically evaluate vision encoder performance, we fix the architecture (Hybrid ResNet-ViT) and data dimensionality (3D) while ablating pretraining strategies. Table 3 compares combinations of three alignment techniques: image-text contrastive learning (ITC), voxel-text alignment (Seg.), and category description alignment (Clas.) across image-text retrieval, segmentation, and classification. Our analysis reveals that multi-granular alignment consistently enhances vision representation quality, with the combination of all three strategies achieving overall optimal performance across tasks. Critically, voxel-text alignment demonstrates superior efficacy, improving image-text retrieval (Average IR@10 +2.75%), diagnosis (AUC +2.09%), and segmentation (Average Dice +8.57%) over the ITC baseline, likely attributable to its spatially grounded supervision. These findings reveal multi-granular alignment, particularly voxel-text supervision, as essential for extracting discriminative features from 3D medical volumes.

4.3 Vision Encoder Architecture Ablation

Building on our optimal multi-granular pretraining strategy and 3D data configuration, we rigorously evaluate three vision encoder architectures: the dominant ViT3D (ViT.), ResNet3D (Res.), and our Hybrid ResNet-ViT (Hyb.). Performance across critical clinical tasks is quantified in Table 4. The hybrid architecture demonstrates superior discriminative capability in diagnosis (ACC +5.29% vs. ResNet3D, +5.38% vs. ViT3D) and segmentation (average DSC +7.40% vs. ResNet3D, +4.52% vs. ViT3D) task, attributing to its synergistic design: convolutional layers capture local spatial features, while self-attention models global context across spatial-depth dimensions. While ViT3D excels on CT-RATE image-text retrieval, Hybrid ResNet-ViT achieves balanced proficiency across diagnosis, segmentation, and retrieval tasks. Consequently, we adopt Hybrid ResNet-ViT for following experiments.

4.4 Efficiency of Cross-Dimensional Learning

Despite 3D CT being the clinical standard, vast repositories of richly annotated 2D slice data coexist in medical practice. We validate whether jointly training on both modalities enhances 3D model performance. Through systematic comparison of 3D-only training (public 3D data) versus integrated 2D+3D training (public 3D and 2D slice datasets), Table 5 shows performance gains from cross-dimensional pretraining. These improvements confirm that cross-dimensional pretraining provides complementary representation learning, effectively bridging the dimensional asymmetry between slices and volumes to strengthen 3D analysis.

4.5 CTInstruct v.s. SOTA MVLMs

In this part, we compare CTInstruct with six SOTA baselines on CT modality, including three 3D generative MVLMs

Method	CT-RATE				Rad-ChestCT			
	AUC	ACC	F1-Score	Precision	AUC	ACC	F1-Score	Precision
CT-Net	50.27±22.75 **	38.18±19.52 **	29.45±16.73 *	20.05±13.23 *	51.42±23.94 **	41.02±19.81 **	39.29±24.76 **	28.50±17.92 **
RadFM	53.78±20.16 **	56.05±23.65	25.24±15.61 **	24.29±22.07	50.49±21.42 **	49.79±20.23 **	33.50±23.90 **	26.06±24.34 **
CT-Chat	30.00±26.55 **	59.66±17.19	15.51±12.00 **	14.95±12.37 **	28.74±28.48 **	60.13±18.30	18.41±10.31 **	19.29±12.13 **
CT-CLIP	56.49±5.39	40.84±17.25 **	32.80±13.69	21.40±10.62 *	52.17±3.50 *	43.80±17.59 **	41.29±22.80	29.47±21.75 *
M3D	54.92±29.02 *	45.14±20.23 **	23.59±9.70 **	20.32±11.37 *	43.09±25.44 **	52.53±18.25 **	26.96±13.41 **	31.85±24.19 *
CTInstruct	58.47±4.63	56.83±17.70	33.80±13.11	24.78±9.77	59.42±0.06	63.03±0.18	44.87±0.22	37.36±0.20

Table 6: Medical diagnosis results on CT-RATE and Rad-ChestCT. AUC, ACC, F1-Score, Precision are reported. Darker grid shading indicates better outcomes. All values are percentages (%) with $p < 0.05$ marked with * and $p < 0.01$ marked with **.

Method	CT-VQA			RP3D-VQA			RP3D-Cap		
	ROUGE-1	BLEU-1	METEOR	ROUGE-1	BLEU-1	METEOR	ROUGE-1	BLEU-1	METEOR
RadFM	15.44±14.08**	11.12±10.38**	14.26±12.83**	25.97±21.77**	14.52±11.21**	16.25±12.40**	16.49 ±7.74*	12.23±7.46*	11.57±5.47*
CT-Chat	58.40±39.26	49.41±40.41**	31.12±28.97**	8.80±7.41**	1.93±1.41**	9.23±8.05**	4.43±3.21**	2.95±2.63**	2.64±1.95**
M3D	17.12±15.24 **	10.15±9.46**	11.60±8.84 **	37.40±33.88*	33.64±30.82**	25.36±21.55	19.51±13.64	14.90±12.80	14.32±13.15
CTInstruct	59.11±34.98	54.01±35.87	55.87±34.82	40.66±34.68	37.36 ±24.63	26.58±21.89	18.65±7.07	15.89±7.63	15.33±5.65

Table 7: Medical VQA results on CT-VQA, RP3D-VQA, and RP3D-Cap. Rouge-1, Blue-1, and METEOR are reported. Darker grid shading means better results. All values are percentages (%) with $p < 0.05$ marked with * and $p < 0.01$ marked with **.

(RadFM (Wu et al. 2023), M3D (Bai et al. 2024), CT-Chat (Hamamci et al. 2024b)), two specialized classification models (CT-CLIP (Hamamci et al. 2024a) and CT-Net (Draelos et al. 2021b)), and one specialized segmentation model (SegVol (Du et al. 2024)).

Diagnosis Results. Table 6 demonstrates CTInstruct's overall superior diagnostic performance on both the CT-RATE and Rad-ChestCT datasets. Notably, it achieves substantial AUC improvements over the strongest baselines: +1.98% compared to the specialized diagnosis model CT-CLIP on CT-RATE and +7.25% on Rad-ChestCT. CTInstruct also surpasses the best generative MVLM baseline by +3.55% and +8.93%, respectively. These results indicate that our multi-granular vision pretraining approach, which synergies CLIP-style alignment with specialized discriminative objectives, yields optimal diagnostic capability. While baseline models like CT-Chat attain high accuracy, their significantly lower AUC scores (30.00% on CT-RATE and 28.74% on Rad-ChestCT) and F1-Scores (15.51% on CT-RATE and 18.41% on Rad-ChestCT) highlight critical limitations in handling imbalanced datasets.

Visual Question Answering Results. As shown in Table 7, CTInstruct achieves significant performance gains on the CT-VQA dataset, notably a +24.75% improvement in METEOR. This substantial METEOR increase particularly highlights CTInstruct's enhanced capability for generating semantically coherent answers. On the web-derived RP3D benchmark, CTInstruct consistently maintains SOTA performance across both subtasks: it achieves gains of +3.26% in ROUGE-1, +3.72% in BLEU-1, and +1.22% in METEOR for RP3D-VQA, while maintaining comparable performance to M3D (a generative MVLM tailored for the RP3D dataset) on RP3D-Caption. This consistent superiority across both medical and web-sourced datasets confirms the effectiveness of our unified training paradigm.

Model	TotalSegmentor	Abdomen1K	CT-ORG
SegVol	44.28	79.06	77.78
M3D	59.70	73.37	81.01
CTInstruct	98.07±4.40	96.39±3.47	93.03±8.16

Table 8: Medical segmentation results. Darker grid shading indicates better outcomes. All values are percentages (%).

Segmentation Results. Table 8 demonstrates substantial segmentation improvements by CTInstruct over baseline methods across all datasets, achieving gains of 38.37% on TotalSegmentor, +17.33% on Abdomen1K, and +12.02% on CT-ORG compared to SOTA baselines. These results confirm that multi-granular vision pretraining significantly enhances MVLMs' performance on fine-grained tasks, highlighting the necessity to advance beyond conventional coarse-grained CLIP-style pretraining paradigm.

5 Conclusion

We present CTInstruct, a generative vision-language model designed for versatile downstream radiology applications, including diagnosis, segmentation, and open generation tasks. We conduct systematic ablation studies to optimize the vision encoder for volumetric imaging, introducing a novel vision encoder design that enables cross-dimensional learning and multi-granular image-text alignment. Besides, we introduce a multi-task instruction tuning framework that unifies diverse radiology tasks under a single generative formulation, incorporating dynamic task-specific tokens to handle structural differences between tasks seamlessly. Experiments show CTInstruct achieves SOTA performance across 8 benchmarks spanning diagnostic, segmentation, and generation tasks, offering insights for data-efficient clinical multimodal systems and future medical multimodal learning.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2022ZD0160702), the Scientific Research Innovation Capability Support Project for Young Faculty (ZYGXQNJSKYCXNLZCXM-I22), and the National Natural Science Foundation of China (No. 24Z031503678).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Andrearczyk, V.; Oreiller, V.; Abobakr, M.; Akhavanallaf, A.; Balermpas, P.; Boughdad, S.; Capriotti, L.; Castelli, J.; Cheze Le Rest, C.; Decazes, P.; et al. 2022. Overview of the HECKTOR challenge at MICCAI 2022: automatic head and neck tumor segmentation and outcome prediction in PET/CT. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, 1–30. Springer.

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature communications*, 13(1): 4128.

Bai, F.; Du, Y.; Huang, T.; Meng, M. Q.-H.; and Zhao, B. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bakr, S.; Gevaert, O.; Echegaray, S.; Ayers, K.; Zhou, M.; Shafiq, M.; Zheng, H.; Benson, J. A.; Zhang, W.; Leung, A. N.; et al. 2018. A radiogenomic dataset of non-small cell lung cancer. *Scientific data*, 5(1): 1–9.

Draelos, R. L.; Dov, D.; Mazurowski, M. A.; Lo, J. Y.; Henao, R.; Rubin, G. D.; and Carin, L. 2021a. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67: 101857.

Draelos, R. L.; Dov, D.; Mazurowski, M. A.; Lo, J. Y.; Henao, R.; Rubin, G. D.; and Carin, L. 2021b. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67: 101857.

Du, Y.; Bai, F.; Huang, T.; and Zhao, B. 2024. Segvol: Universal and interactive volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 37: 110746–110783.

Fan, Z.; Du, S.; Hu, S.; Wang, P.; Shen, L.; Zhang, Y.; Tao, D.; and Wang, Y. 2025a. Combatting dimensional collapse in LLM pre-training data via submodular file selection. In *The Thirteenth International Conference on Learning Representations*.

Fan, Z.; Liang, C.; Wu, C.; Zhang, Y.; Wang, Y.; and Xie, W. 2025b. ChestX-Reasoner: Advancing Radiology Foundation Models with Reasoning through Step-by-Step Verification. *arXiv preprint arXiv:2504.20930*.

Gatidis, S.; Hepp, T.; Früh, M.; La Fougère, C.; Nikolaou, K.; Pfannenberg, C.; Schölkopf, B.; Küstner, T.; Cyran, C.; and Rubin, D. 2022. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Scientific Data*, 9(1): 601.

Hamamci, I. E.; Er, S.; Almas, F.; Simsek, A. G.; Esirgun, S. N.; Dogan, I.; Dasdelen, M. F.; Durugol, O. F.; Wittmann, B.; Amiranashvili, T.; et al. 2024a. Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. *arXiv preprint arXiv:2403.17834*.

Hamamci, I. E.; Er, S.; Almas, F.; Simsek, A. G.; Esirgun, S. N.; Dogan, I.; Dasdelen, M. F.; Wittmann, B.; Simsar, E.; Simsar, M.; et al. 2024b. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*.

He, Y.; Yang, G.; Yang, J.; Ge, R.; Kong, Y.; Zhu, X.; Zhang, S.; Shao, P.; Shu, H.; Dillenseger, J.-L.; et al. 2021. Meta grayscale adaptive network for 3D integrated renal structures segmentation. *Medical image analysis*, 71: 102055.

Heller, N.; Isensee, F.; Trofimova, D.; Tejpaul, R.; Zhao, Z.; Chen, H.; Wang, L.; Golts, A.; Khapun, D.; Shats, D.; et al. 2023. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*.

Huang, S.-C.; Huo, Z.; Steinberg, E.; Chiang, C.-C.; Lungren, M. P.; Langlotz, C. P.; Yeung, S.; Shah, N. H.; and Fries, J. A. 2023. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. *arXiv preprint arXiv:2311.10798*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jaus, A.; Seibold, C.; Hermann, K.; Walter, A.; Giske, K.; Haubold, J.; Kleesiek, J.; and Stiefelhagen, R. 2023. Towards unifying anatomy segmentation: automated generation of a full-body CT dataset via knowledge aggregation and anatomical guidelines. *arXiv preprint arXiv:2307.13375*.

Ji, Y.; Bai, H.; Ge, C.; Yang, J.; Zhu, Y.; Zhang, R.; Li, Z.; Zhanng, L.; Ma, W.; Wan, X.; et al. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35: 36722–36732.

Jun, M.; Cheng, G.; Yixin, W.; Xingle, A.; Jiantao, G.; Ziqi, Y.; Minqing, Z.; Xin, L.; Xueyuan, D.; Shucheng, C.; et al. 2020. COVID-19 CT lung and infection segmentation dataset. *(No Title)*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Lambert, Z.; Petitjean, C.; Dubray, B.; and Kuan, S. 2020. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. Ieee.

Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, 12. Munich, Germany.

Lee, R. K.-W.; Cao, R.; Fan, Z.; Jiang, J.; and Chong, W.-H. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, 5138–5147.

Lei, J.; Dai, L.; Jiang, H.; Wu, C.; Zhang, X.; Zhang, Y.; Yao, J.; Xie, W.; Zhang, Y.; Li, Y.; et al. 2025. Unibrain: Universal brain mri diagnosis with hierarchical knowledge-enhanced pre-training. *Computerized Medical Imaging and Graphics*, 122: 102516.

Lei, J.; Zhang, X.; Wu, C.; Dai, L.; Zhang, Y.; Zhang, Y.; Wang, Y.; Xie, W.; and Li, Y. 2024. Autorg-brain: Grounded report generation for brain mri. *arXiv preprint arXiv:2407.16684*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, P.; Han, H.; Du, Y.; Zhu, H.; Li, Y.; Gu, F.; Xiao, H.; Li, J.; Zhao, C.; Xiao, L.; et al. 2021. Deep learning to segment pelvic bones: large-scale CT datasets and baseline models. *International Journal of Computer Assisted Radiology and Surgery*, 16(5): 749–756.

Luo, G.; Wang, K.; Liu, J.; Li, S.; Liang, X.; Li, X.; Gan, S.; Wang, W.; Dong, S.; Wang, W.; et al. 2023. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge. *arXiv preprint arXiv:2304.03708*.

Luo, X.; Liao, W.; Xiao, J.; Chen, J.; Song, T.; Zhang, X.; Li, K.; Metaxas, D. N.; Wang, G.; and Zhang, S. 2022. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. *Medical Image Analysis*, 82: 102642.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024a. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Ma, J.; Zhang, Y.; Gu, S.; Ge, C.; Mae, S.; Young, A.; Zhu, C.; Yang, X.; Meng, K.; Huang, Z.; et al. 2024b. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the FLARE22 challenge. *The Lancet Digital Health*, 6(11): e815–e826.

Ma, J.; Zhang, Y.; Gu, S.; Zhu, C.; Ge, C.; Zhang, Y.; An, X.; Wang, C.; Wang, Q.; Liu, X.; et al. 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6695–6714.

Masoudi, M.; Pourreza, H.-R.; Saadatmand-Tarzjan, M.; Eftekhari, N.; Zargar, F. S.; and Rad, M. P. 2018. A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. *Scientific data*, 5(1): 1–9.

Mei, X.; Liu, Z.; Robson, P. M.; Marinelli, B.; Huang, M.; Doshi, A.; Jacobi, A.; Cao, C.; Link, K. E.; Yang, T.; Wang, Y.; Greenspan, H.; Deyer, T.; Fayad, Z. A.; and Yang, Y. 0. RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, 0(ja): e210315.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, 353–367. PMLR.

Nguyen, D.; Kay, F.; Tan, J.; Yan, Y.; Ng, Y. S.; Iyengar, P.; Peshock, R.; and Jiang, S. 2021. Deep learning–based COVID-19 pneumonia classification using chest CT images: model generalizability. *Frontiers in Artificial Intelligence*, 4: 694875.

Pedrosa, J.; Aresta, G.; Ferreira, C.; Atwal, G.; Phoulady, H. A.; Chen, X.; Chen, R.; Li, J.; Wang, L.; Galdran, A.; et al. 2021. LNDb challenge on automatic lung cancer patient management. *Medical image analysis*, 70: 102027.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Radl, L.; Jin, Y.; Pepe, A.; Li, J.; Gsaxner, C.; Zhao, F.-h.; and Egger, J. 2022. AVT: Multicenter aortic vessel tree CTA dataset collection with ground truth segmentation masks. *Data in brief*, 40: 107801.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Remy, F.; Demuynck, K.; and Demeester, T. 2022. Biolord: Learning ontological representations from definitions (for biomedical concepts and their textual descriptions). *arXiv preprint arXiv:2210.11892*.

Rister, B.; Yi, D.; Shivakumar, K.; Nobashi, T.; and Rubin, D. L. 2020. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1): 381.

Roth, H. R.; Lu, L.; Farag, A.; Shin, H.-C.; Liu, J.; Turkbey, E. B.; and Summers, R. M. 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, 556–564. Springer.

Setio, A. A. A.; Traverso, A.; De Bel, T.; Berens, M. S.; Van Den Bogaard, C.; Cerello, P.; Chen, H.; Dou, Q.; Fantacci, M. E.; Geurts, B.; et al. 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical image analysis*, 42: 1–13.

Soares, E.; Angelov, P.; Biaso, S.; Froes, M. H.; and Abe, D. K. 2020. SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification. *MedRxiv*, 2020–04.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical AI. *Nejm Ai*, 1(3): AIoa2300138.

Wasserthal, J.; Breit, H.-C.; Meyer, M. T.; Pradella, M.; Hinck, D.; Sauter, A. W.; Heye, T.; Boll, D. T.; Cyriac, J.; Yang, S.; et al. 2023. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5): e230024.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *arXiv preprint arXiv:2308.02463*.

Xue, L.; Shu, M.; Awadalla, A.; Wang, J.; Yan, A.; Purushwalkam, S.; Zhou, H.; Prabhu, V.; Dai, Y.; Ryoo, M. S.; et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Yang, J.; Shi, R.; Jin, L.; Huang, X.; Kuang, K.; Wei, D.; Gu, S.; Liu, J.; Liu, P.; Chai, Z.; et al. 2025. Deep rib fracture instance segmentation and classification from ct on the ribfrac challenge. *IEEE Transactions on Medical Imaging*.

Zhang, M.; Wu, Y.; Zhang, H.; Qin, Y.; Zheng, H.; Tang, W.; Arnold, C.; Pei, C.; Yu, P.; Nan, Y.; et al. 2023. Multi-site, multi-domain airway tree modeling. *Medical image analysis*, 90: 102957.

Zhao, J.; Zhang, Y.; He, X.; and Xie, P. 2020. COVID-CT-Dataset: a CT scan dataset about COVID-19. *arXiv preprint arXiv:2003.13865*.

Zhao, Z.; Zhang, Y.; Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*.

Zheng, Q.; Zhao, W.; Wu, C.; Zhang, X.; Dai, L.; Guan, H.; Li, Y.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. Large-scale long-tailed disease diagnosis on radiology images. *Nature Communications*, 15(1): 10147.

Zhuang, X.; and Shen, J. 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical image analysis*, 31: 77–87.