Medical Vision Generalist (MVG): Unifying Medical Imaging Tasks in Context

Anonymous authors
Paper under double-blind review

Abstract

This study presents Medical Vision Generalist (MVG), a vision-centric generalist model capable of handling various medical imaging tasks—such as cross-modal synthesis, image segmentation, denoising, and inpainting—within a unified single image-to-image generation framework. Specifically, MVG employs an in-context generation strategy that standardizes the handling of inputs and outputs as images. By treating these tasks as an image generation process conditioned on prompt image-label pairs and input images, this approach enables a flexible unification of various tasks, even those spanning different modalities and datasets. To capitalize on both local and global context, we design a hybrid method combining masked image modeling with autoregressive training for conditional image generation. This hybrid approach yields the most robust performance across all involved medical imaging tasks. To rigorously evaluate MVG's capabilities, we curated the first comprehensive generalist medical vision benchmark, comprising 13 datasets, over 2 million training images, and spanning four imaging modalities (CT, MRI, X-ray, and micro-ultrasound). Our results consistently establish MVG's superior performance, outperforming existing vision generalists, such as Painter and LVM, and in some cases, matching the efficacy of specialized task-specific models. Furthermore, MVG exhibits strong scalability, with its performance demonstrably improving when trained on a more diverse set of tasks, and can be effectively adapted to unseen datasets with only minimal task-specific samples. Code will be made available.

1 Introduction

The precise interpretation of medical images is imperative for timely disease detection, diagnosis, and treatment Cheng et al. (2022b); De Fauw et al. (2018). Deep-learning based models have emerged as powerful tools in medical image analysis, tackling various challenges spanning from segmenting specific anatomical structures (Ji et al., 2022; Luo et al., 2021; Fu et al., 2021), localizing single organ diseases (Zhu et al., 2019; Zhao et al., 2021; Huo et al., 2020; Cheng et al., 2022a; Ardila et al., 2019; Kim et al., 2022; Heller et al., 2021), to cross-modality image synthesis on brain MRI (Xie et al., 2023a; Li et al., 2023; Dayarathna et al., 2023; Zhu et al., 2023). However, these models, often referred to as specialist models, are typically customized for specific tasks, modalities, or anatomical regions. While this specialization often results in exceptional performance in certain contexts, it can lead to a severe performance drop when applied to new tasks or when tasked with training multi-domain data.

To address this challenge, recently, there has been a partial shift of research focus in developing generalist medical AI models (Moor et al., 2023; Tu et al., 2023), which necessitate only a single training phase but are capable of wide application across a diverse array of medical tasks. Specifically, these generalist frameworks unify input and output spaces, allowing straightforward adaptation to various tasks through user-provided prompts. While existing generalist medical AI models like MedSAM have demonstrated impressive performance (Ma et al., 2024; Zhang et al., 2023; Butoi et al., 2023), their applicability in medical visual tasks remains limited (*e.g.*, to segmentation tasks only). A unified, truly generalist vision model capable of addressing a vast array of medical imaging tasks remains a critical missing piece in the current medical research landscape.

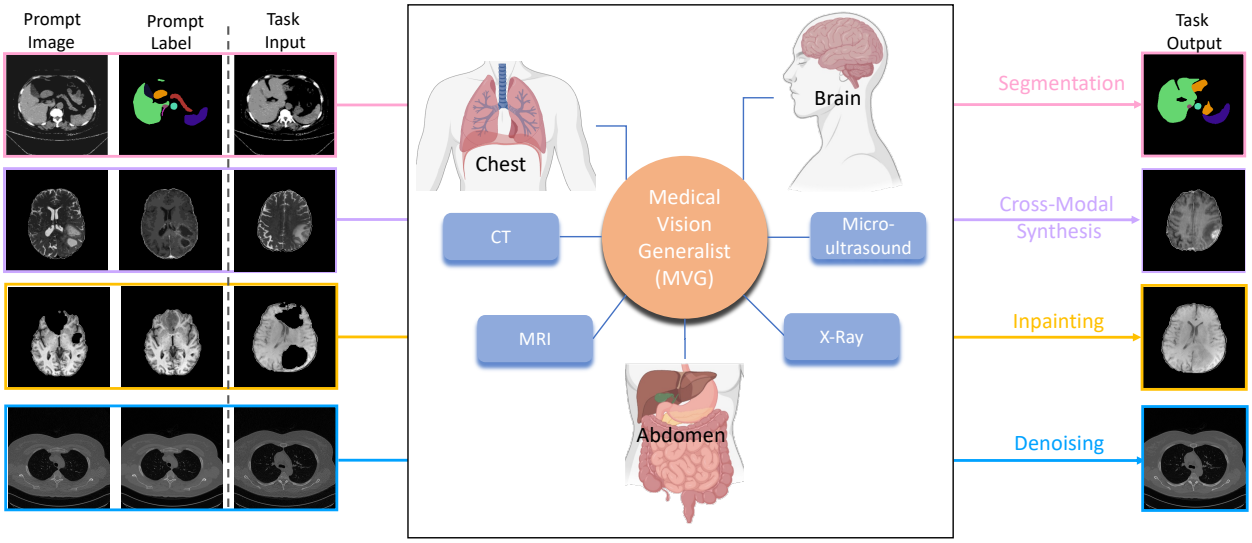

Figure 1: **Medical Vision Generalist** enables a single model to be capable of performing **four** types of medical vision tasks on images in **four** medical imaging modalities of multiple body regions.

Inspired by the remarkable success of in-context learning in natural language processing (Brown et al., 2020a; OpenAI, 2023) and computer vision (Bai et al., 2023; Wang et al., 2023b;a), we propose the Medical Vision Generalist (MVG)—a proof-of-concept demonstrating the feasibility of a vision-centric generalist model capable of performing diverse imaging tasks in the medical domain without the need for text-based guidance. Specifically, MVG leverages an in-context learning framework to unify a set of medical imaging tasks, including cross-modal synthesis, denoising, segmentation, and inpainting across modalities like CT, MRI, X-ray, and Micro-ultrasound. In contrast to prior task- and data-specific medical AI models, MVG offers adaptability to new data with minimal labeled samples, eliminating the need for retraining. To achieve this, MVG first standardizes the input/output space using in-context coloring, which maps various tasks into a single-channel coloring scheme. This removes the need for task-specific heads, thus regulating the model to learn exclusively from prompts. Subsequently, tasks are unified through conditional image generation, where MVG generates the output conditioned on both the task prompt and a sample image.

To capture both local and global context, we devise a hybrid strategy that combines masked image modeling and autoregressive training for conditional image generation. The former involves concatenating prompt images, labels, task inputs, and labels, followed by random masking; the latter constructs prompt image-label pairs, task inputs, and labels as long visual sentences. During inference, MVG conditions predictions on the prompts selected from locations closely matching the task images, ensuring contextual relevance and guidance that enhances output quality and consistency.

Furthermore, we have curated the first unified medical imaging benchmark, encompassing 13 datasets spanning a range of human anatomies (*e.g.*, abdomen, pelvis, brain, chest) and modalities (*e.g.*, CT, MRI, X-ray, micro-ultrasound). This new benchmark enables a comprehensive assessment of our MVG models. Experimental results demonstrate the effectiveness of our MVG in performing various medical vision tasks with only one model. As illustrated in Figure 2, our MVG outperforms the previous generalist models by a large margin. For instance, our MVG achieves 0.735 mIoU on all segmentation tasks and outperforms the previous best vision generalist by 0.123 mIoU. Furthermore, our MVG demonstrates two intriguing properties: 1) it scales well with multiple tasks and datasets, suggesting its potential to excel further as diverse datasets continue to emerge; and 2) it can efficiently generalize to new datasets, with only a few specific examples needed for each task.

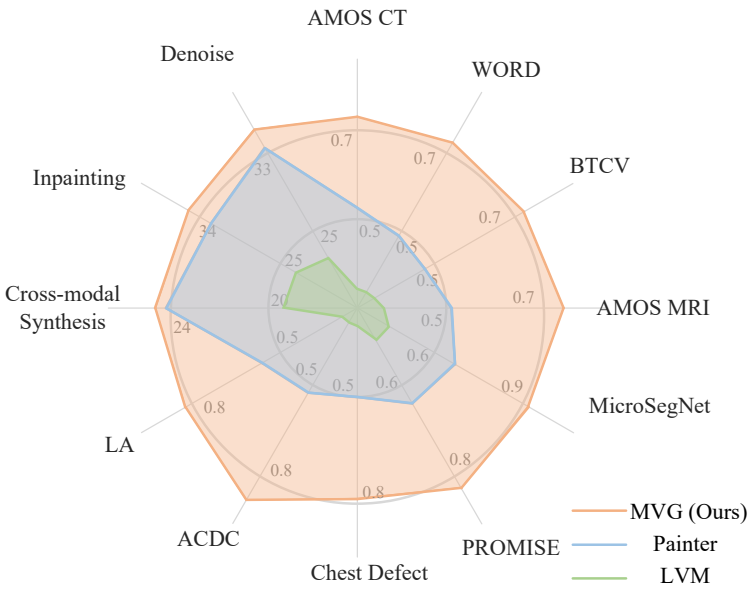

Figure 2: **Comparison with other generalists.** Our model achieves state-of-the-art performance on all involved medical vision tasks of five types.

2 Related Work

Medical Image Analysis In the field of medical image analysis, there have been key developments in deep-learning models for image segmentation. As the earliest success in this line of work, U-Net (Ronneberger et al., 2015) uses an encoder-decoder architecture with skip-connection, revealing the great potential of deep networks. Following the line, nnUnet (Isensee et al., 2021) further improves the model architecture and introduces bags of tricks, building a well-engineered general segmentation model. TransUnet (Chen et al., 2021a) proposes to use pre-trained ViT for better feature extraction. Recent efforts in medical image analysis have produced remarkable models capable of performing a variety of tasks. Notable works include "One model to rule them all" (Zhao et al., 2023), MedSAM (Ma et al., 2024), and UniverSeg (Butoi et al., 2023), which are designed to tackle unified medical segmentation tasks. UniverSeg adapts UNet (Ronneberger et al., 2015) to intake in-context samples to segment new data and tasks without further training. Besides, biomedGPT (Zhang et al., 2023) proposes a unified generative model for bio-medical vision-language tasks. In this paper, we propose a novel paradigm to build a generalist model, which is capable of handling various medical vision tasks, including segmentation, inpainting, cross-modal synthesis, and denoising.

Universal Models and In-Context Learning The advent of the universal Transformer architecture and its success in generative pretraining has inspired the development of universal models that tackle a wide range of computer vision tasks (Chen et al., 2021b; 2022; Lu et al., 2022; Wang et al., 2022; Bai et al., 2023; Wang et al., 2023a). In-context learning is a novel few-shot learning paradigm that emerged in large language models and was first proposed by GPT-3 (Brown et al., 2020b). Specifically, in-context learning enables one model to perform different tasks with only in-context examples as prompts. While the prompts for language models are mostly defined as a few sentences, in-context learning in other domains is still in an early exploration stage. As one of the earliest works, Flamingo (Alayrac et al., 2022) extends the modality of in-context learning with language instructions and sequences of images and videos. Perceiver-IO (Jaegle et al., 2021) uses the Transformer architecture for a general-purpose model that handles data from arbitrary settings like natural language, visual understanding, multi-modal reasoning, and StarCraft II. AD (Laskin et al., 2022) introduces in-context learning to reinforcement learning with algorithm distillation. DPT (Lee et al., 2024) provides a sample-efficient RL algorithm with strong in-context decision-making. In this paper, we use a sequence of paired medical images to build a vision model with in-context learning ability, unifying 13 medical tasks as a generation task.

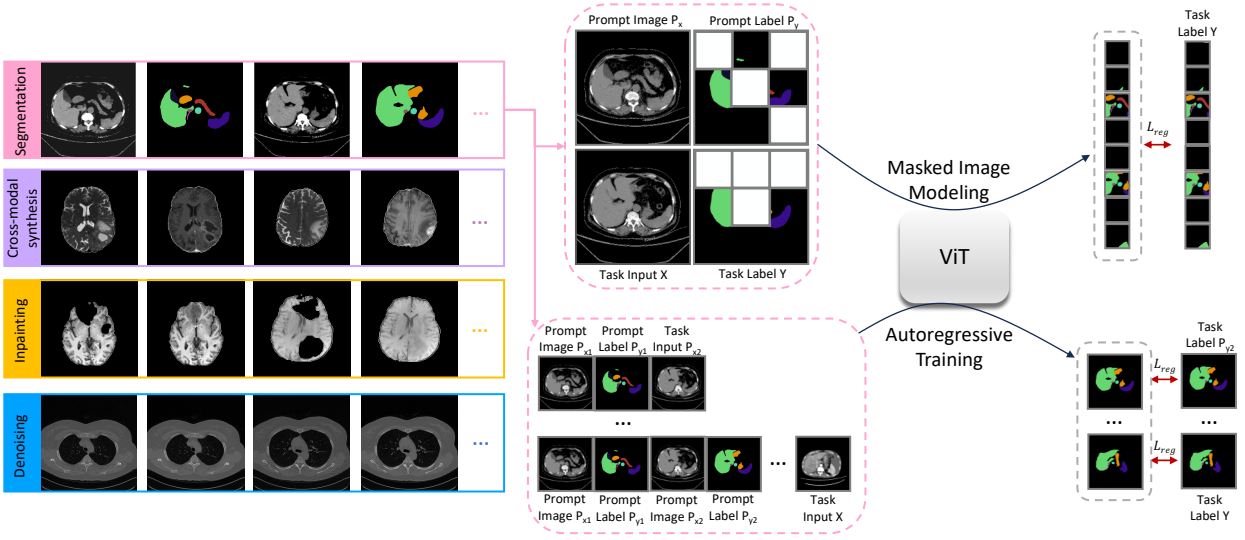

Figure 3: **Method overview. Left:** Four types of medical tasks (*i.e.*, segmentation, cross-modal synthesis, inpainting, and denoising) are unified as a universal image-to-image generation task with in-context learning. **Right:** We adopt mask image modeling and auto-regressive training for in-context generation.

3 Method

Unlike previous medical AI models, which are specific to one or a few predefined imaging tasks and produce a predetermined set of outputs, the proposed MVG aims to offer unprecedented flexibility across tasks, modalities, and datasets. The key idea is to unify medical imaging tasks, such as cross-modal synthesis, image segmentation, denoising, and inpainting, within an image-to-image generation framework.

3.1 Tasks

Our MVG is designed to address various medical imaging tasks, with a particular emphasis in this study on segmentation, cross-modal synthesis, inpainting, and denoising tasks for which well-represented public datasets are available. However, it is crucial to note that its design should be widely applicable to any single image-to-image generation task.

Segmentation. Medical image segmentation, including CT, MRI, X-ray, and Micro-ultrasound segmentation, involves dividing an image obtained from these modalities into distinct segments to isolate regions of interest, such as organs or abnormalities. The input space for these tasks typically consists of images from CT, MRI, X-ray, or Micro-ultrasound scans. The output space is represented by a mask, where each value (excluding the background) in the mask corresponds to a different class or type of object, such as a liver or kidney.

Cross-modal synthesis. Cross-modal synthesis aims to generate images in one modality from images of another modality for the same subject, aiding in visualization and facilitating multi-modal medical image analysis. The input space and output space are different medical imaging modalities.

Brain image inpainting. In the context of brain image processing, inpainting refers to the process of synthesizing healthy brain tissue in regions affected by glioma, a type of brain tumor (Kofler et al., 2023). Inpainting allows professionals to effectively utilize non-standard imaging protocols and directly apply brain parcellation tools to facilitate treatment planning. The input space is the corrupted brain MRI and the output space is the corresponding brain MRI restoring the affected regions to a normal state.

Denoising. Denoising aims to reconstruct full-dose CT images from low-dose CT images, allowing for reduced radiation doses during CT scans while preserving diagnostic image quality. The input space is the

scanned CT image with low-dose radiation, while the output space is the corresponding image with full-dose radiation.

3.2 Unifying the Input/Output Space

Assume an input image is denoted as $\mathbf{x} \in \mathcal{R}^{H \times W}$, the output could be a segmentation map, a synthesized brain image in the target modality, a restored normal brain MRI, or a full-dose CT image of the same size. To unify the output space of images across tasks, our MVG adopts a strategy beyond task-specific heads: mapping different tasks into a single-channel coloring scheme, inspired by (Wang et al., 2023a;b). Specifically, we explore three different in-context coloring methods for segmentation that circumvent reliance on label values, including binary, pre-defined, and random colorization.

Binary colorization. We break down the problem of segmenting multiple classes into individual binary segmentation tasks, each focusing on separating one class from the background. Specifically, if a segmentation mask contains N_k foreground classes, we simply split it to N_k binary masks. However, this requires multiple inferences when an image contains more than one foreground class.

Pre-defined colorization. In this approach, we allocate a predetermined unique color to each segmentation mask derived from diverse datasets. Suppose there are K segmentation datasets, with each dataset containing N_k classes. Consequently, the n_{th} class of the k_{th} dataset is assigned the value of $\sum_{i=1}^{k-1} i * N_i + n$. Note that different tasks may involve classes with identical semantics; for example, both the AMOS segmentation dataset (Ji et al., 2022) and the Synapse dataset (Landman et al., 2015) include the class "Liver". However, distinct colors are assigned to the same class across different tasks.

Random colorization. The use of pre-defined colors may restrict the adaptability and efficacy of MVG, as they can cause the model to focus on learning tasks based on the color of the prompt rather than the contextual information (Wang et al., 2023b). To address this limitation, we build a set of colors and randomly sample colors for different semantics in one iteration but the same semantic in the prompt label and task label share the same color.

Except for medical image segmentation, the outputs of all other tasks in this study do not involve categorical values that need to be predicted. Therefore, we do not apply coloring for these tasks.

3.3 Task Unification via Conditional Image Generation

After standardizing the input and output space for all tasks as images of identical sizes, we construct the training input, including 1) the task prompt consisting of paired prompt images and prompt labels, and 2) the task input and its associated label. We then unify various medical imaging tasks within a conditional image-to-image generation framework using the task prompt as task specification. All the tasks are unified to generate the task label $Y \in \mathcal{R}^{H \times W}$ based on the condition including the task image $X \in \mathcal{R}^{H \times W}$, the prompt image $P_x \in \mathcal{R}^{H \times W}$, and the prompt label $P_y \in \mathcal{R}^{H \times W}$. Specifically, we use two conditional image generation frameworks: masked image modeling (He et al., 2022) and autoregressive training (Chen et al., 2020).

Architecture Selection Following the same setting in (He et al., 2022; Hua et al., 2022), we take vanilla ViT (Dosovitskiy et al., 2020) as an encoder including a patch embedding layer and several Transformer blocks. The decoder is a simple prediction head with two convolution layers and takes four feature maps (Li et al., 2022) from ViT as input.

Mask Image Modeling During training, we form a square image by concatenating the prompt image (upper left) with its corresponding label (upper right), as well as the task image (lower left) with its associated label (lower right), as illustrated in Figure 3. We perform random masking on the square image and train ViT to reconstruct the masked region (Wang et al., 2023a):

$$p(x) = \prod_{i=1}^{M} p(x_i | x_{x \notin x_M}, \theta). \tag{1}$$

Region	Dataset	Modality	#Training	#Testing	Task
Abdomen	AMOS (Ji et al., 2022)	CT	240	120	Segmentation
Abdomen	WORD (Luo et al., 2021)	CT	100	20	Segmentation
Abdomen	BTCV (Iglesias & Sabuncu, 2015)	CT	21	9	Segmentation
Abdomen	AMOS (Ji et al., 2022)	MRI	60	50	Segmentation
Pelvis	MicroSegNet (Jiang et al., 2024)	Micro-US	55	20	Segmentation
Pelvis	PROMISE (Litjens et al., 2014)	MRI	50	30	Segmentation
Brain	BraTS-GLI (Kazerooni et al., 2023)	MRI	1251	219	Cross-modal synthesis
Brain	BraTS-Local (Kazerooni et al., 2023)	MRI	1000	251	Inpainting
Chest	Low dose (McCollough et al., 2021)	CT	200	59	Denoising
Chest	Defect Detection (Candemir & Antani, 2019)	Xray	15	6	Segmentation
Chest	ACDC (Bernard et al., 2018)	MRI	100	50	Segmentation
Chest	LA (Chen et al., 2019)	MRI	81	20	Segmentation
Whole body	Deeplesion (Yan et al., 2018)	CT	25000	7120	Detection

Table 1: **Datasets overview.** Our MVG is trained and evaluated on 13 different datasets covering four major human body regions (*i.e.*, Abdomen, Pelvis, Brain, Chest). #Training/Testing refers to the number of samples for training and testing.

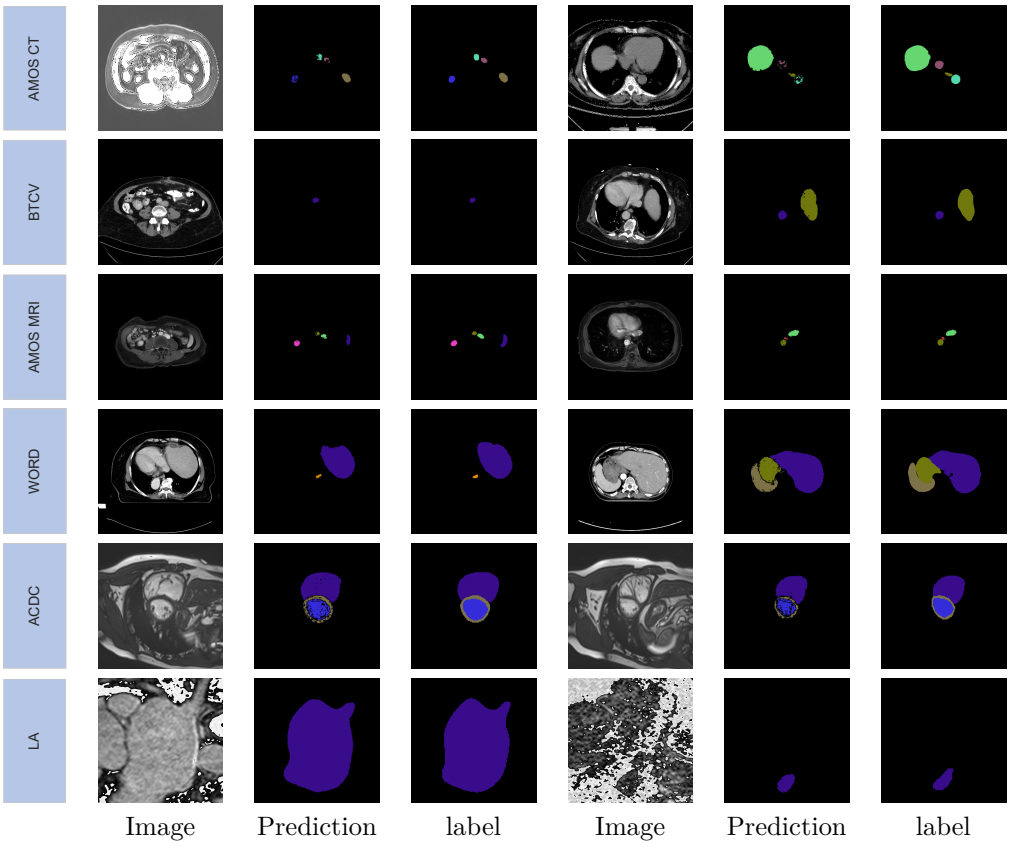

Figure 4: **Qualitative evaluation of segmentation.** MVG shows strong capabilities on various segmentation tasks covering multiple modalities and body regions.

where x_M is the mask region, $x_{x \notin x_M}$ is the visible region, and θ denotes the model parameters. However, in practice, we observed that mask image modeling yields unsatisfactory results for medical image segmentation. We hypothesize that this may be attributed to the masking strategy's potential to compromise the preservation of global contextual information within individual images, such as the interplay among various abdominal organs. Furthermore, for small organs like the pancreas and the gallbladder, this masking approach can render them nearly invisible in prompts and makes prompts provide no in-context information

about these organs. In contrast, for tasks like inpainting and denoising, masked image modeling excels, as these tasks prioritize refining local details over preserving global contextual information. To ensure the efficacy on medical image segmentation, we introduce an additional *auto-regressive* training, which preserves the global context within individual images, as shown below.

Auto-Regressive Training In auto-regressive training, each image, including paired prompt images, prompt labels, task inputs, and associated labels, is treated as a single element in a sequential data structure. The model is fed with a partial sequence and trained to predict the next image in the sequence conditioning on the preceding ones.

Mathematically, let $P_{x_1}, P_{y_1}, ..., P_{x_n}, P_{y_n}, X, Y$ denote $n+1$ pairs of images and labels. The first n pairs serve as the task prompt, and the model learns to predict the task output Y given the task input X and the prompt. This process iterates through each pair in the sequence. For each iteration, auto-regressive training is conducted with supervision solely on prompt labels and the task label:

$$\begin{aligned} S &= [S_1, S_2, ..., S_{2n-1}, S_{2n}, S_{2n+1}, S_{2n+2}] \\ &= [P_{x_1}, P_{y_1}, ..., P_{x_n}, P_{y_n}, X, Y], \\ p(x) &= \prod_{i=1}^{n+1} p(S_{2i}|S_1, ..., S_{2i-1}, \theta). \end{aligned} \tag{2}$$

Loss Function Any regression loss function like $\boldsymbol{l}_1$ or $\boldsymbol{l}_2$ can serve as the loss function of our MVG. Different from the $\boldsymbol{l}_2$ loss function in masked image modeling (He et al., 2022), we find the smooth $\boldsymbol{l}_1$ performs best for MVG.

Inference We first construct a sequence $S = [P_x, P_y, X, \hat{Y}]$, where prompts, the task image, and the desired output are concatenated together. MVG leverages the task prompt, composed of the prompt image P_x and label P_y, for task specification, subsequently generating predictions by conditioning on both the task input X and the task prompt. Since the task prompt is formulated as images, MVG demonstrates versatility in defining imaging tasks, capable of handling data sourced from diverse scanning machines, procedures, settings, or populations. For instance, if P_x and P_y represent an image and label extracted from the AMOS CT training set, respectively, MVG performs multi-organ segmentation on the image X derived from the AMOS CT testing set, maintaining consistency within the dataset setting and guided by the provided context.

Different task prompts can yield varying results. In this study, we address this variability by selecting the prompt image from a location that closely matches the task image. Given the instance X_{TE} which has N_{TE} slices from the testing set, we randomly choose an instance X_{TR} which has N_{TR} slices and the corresponding label Y_{TR} from the training set. For the n_{th} slice of X_{TE}, we also choose the the floor($\frac{n_{th} * N_{TR}}{N_{TE}}$) slice as the prompt.

4 Experiment

4.1 Implementation Details

Data As shown in Table 1, our model is developed on 13 different datasets including 2.5M training images covering four major human body regions (*i.e.*, Abdomen, Pelvis, Brain, Chest). Following the standard preprocessing strategy, we apply a windowing range of [-100, 200] to all involved CT scans for better contrast. Input images are firstly resized to 512 $\times$512 and then randomly cropped with a size of 448 $\times$ 448. To evaluate the generalization of MVG, we choose MSD (Antonelli et al., 2022), a multi-organ segmentation dataset as an out-of-distribution dataset.

Training details AdamW optimizer is used with a weight decay of 0.05. The peak learning is set to $1e^{-3}$ with a cosine learning rate scheduler. We train our model 100 epochs with 5 warm-up epochs. We only adopt the random crop as the data augmentation. The sampling weight of segmentation tasks is 0.5 while the rest of the tasks share 0.5. We use 8 A5000 GPUs to train our models. We use 1 in-context sample for both training and inference.

Method	AMOS CT	WORD	BTCV	AMOS MRI	MicroSegNet	PROMISE	Chest Defect	ACDC	LA
Specialists									
ResNet-18	0.55	0.50	0.51	0.53	0.67	0.75	0.62	0.69	0.68
UNet	0.81	0.83	0.82	0.81	0.90	0.91	0.89	0.86	0.83
VNet	0.70	0.75	0.72	0.73	0.90	0.89	0.86	0.87	0.84
TranUNet	0.80	0.82	0.84	0.82	0.94	0.90	0.88	0.88	0.84
nnUNet	0.87	0.90	0.91	0.88	0.97	0.93	0.90	0.90	0.89
Generalists									
UniverSeg*	0.20	0.29	0.37	0.25	0.71	0.55	0.55	0.54	0.57
Painter	0.52	0.48	0.45	0.51	0.69	0.68	0.50	0.52	0.55
LVM	0.12	0.14	0.10	0.15	0.36	0.30	0.10	0.12	0.13
SegGPT	0.66	0.66	0.65	0.71	0.88	0.75	0.68	0.70	0.71
MVG	0.73	0.74	0.73	0.74	0.91	0.85	0.79	0.85	0.81

Table 2: **Quantitative evaluation in segmentation tasks.** Compared to other generalists, our method achieves state-of-the-art performance with solid improvements. *: We inference the official weights with 64 in-context samples from the training set.

Method	Cross-modal synthesis			Inpainting			Denoise		
	MAE	PSNR	SSIM	MAE	PSNR	SSIM	MAE	PSNR	SSIM
Specialists									
ResNet-18	0.026	20.984	0.860	0.008	30.981	0.959	0.022	30.519	0.709
Pix2Pix	0.018	24.311	0.899	0.008	34.891	0.982	0.020	33.011	0.730
TranUNet	0.016	25.541	0.938	0.005	35.561	0.989	0.016	33.999	0.761
Generalists									
Painter	0.021	24.031	0.920	0.006	33.595	0.978	0.020	33.104	0.721
MVG	0.019	24.721	0.929	0.006	34.521	0.981	0.018	33.521	0.731

Table 3: **Quantitative comparison with other tasks.** Our model shows strong capabilities in the tasks of cross-modal synthesis, impainting, and denoising.

Training Objective In practice, for *all tasks except segmentation*, we perform 90% training iterations with *mask image modeling* and 10% training iterations with *auto-regressive* training. For *segmentation* tasks, we perform 100% training iterations with *auto-regressive* training.

Evaluation We use mean IoU (mIoU) as the evaluation metric for segmentation. For cross-modal synthesis, inpainting, and denoising, we use mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM) as the evaluation metric.

4.2 A Generalist to 13 Medical Tasks

Baselines Our generalist baselines include LVM (Bai et al., 2023) and Painter (Wang et al., 2023a), which are trained on our benchmark. UniverSeg (Butoi et al., 2023) is used as a segmentation generalist baseline. While our specialist baselines include ResNet-18 (He et al., 2016) with a two-layer MLP decoder, UNet (Ronneberger et al., 2015), VNet (Milletari et al., 2016), TransUNet (Chen et al., 2021a), and nnUNet (Isensee et al., 2021). For synthesis tasks, we involve Pix2Pix (Isola et al., 2017) as an additional baseline.

Quantitative evaluation In Table 2, we compare our method with the latest vision generalist models (Bai et al., 2023; Wang et al., 2023a) across a range of segmentation tasks. To ensure a fair comparison, all results were evaluated using the same testing split. Our MVG achieves the best performance of 0.79 mIoU among all the generalists. Specifically, our MVG outperforms Painter (Wang et al., 2023a) by 0.24 mIoU, LVM (Bai et al., 2023) by 0.62 mIoU on average, SegGPT by 0.09 mIoU on average, and UniverSeg (Butoi et al., 2023) by 0.35 mIoU. All the generalists only require one model to perform these different tasks. UniverSeg

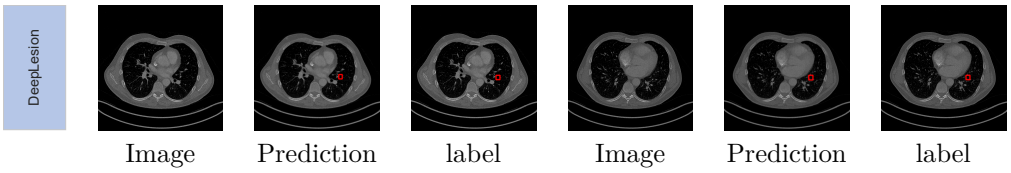

Figure 5: **Qualitative evaluation on DeepLesion detection dataset.**

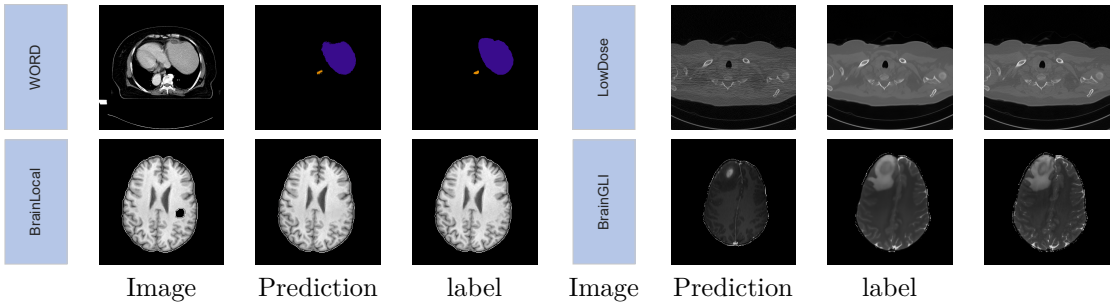

Figure 6: **Qualitative evaluation of four tasks.** Segmentation (1st row), denoising (2nd row), inpainting (4th row) and cross-modal synthesis (3th row).

is trained with up to 64 in-context samples, yet still yields inferior performance to our method which only relies on one in-context sample. We provide the visualization results in Figure 4

At the same time, specialist models like UNet (Ronneberger et al., 2015), TranUNet (Chen et al., 2021a), and nnUNet (Isensee et al., 2021), which need to train different models for different tasks, still hold the edge in performance.

We report the image synthesis results in Table 3. we present a detailed quantitative comparison of our method with the latest generalist and specialist models across various tasks including cross-modal synthesis, inpainting, and denoising. Our MVG demonstrates strong capabilities and achieves competitive performance, particularly in the generalist category. For instance, in the task of cross-modal synthesis, MVG shows an improvement over Painter in all metrics: a lower MAE by 0.002, a higher PSNR by 0.69, and a better SSIM by 0.009 over the best vision generalist.

Qualitative evaluation To provide a more intuitive observation of our MVG, we provide the visualization of different tasks in Figure 6.

Other tasks We show that beyond segmentation, our MVG can handle more discriminative tasks such as object detection. Unlike standard object detection outputs, we form the output space as the original image with the lesion's bounding box overlaid to indicate its location. Specifically, we also add a large-scale lesion detection dataset, DeepLesion (Yan et al., 2018), when training MVG, which aims to identify and localize abnormalities in Chest CT images. These abnormalities include tumors, cysts, and other pathological changes within body tissues, organs, or bones. Our results, illustrated in Figure 5, demonstrate the efficacy of MVG in this context. This suggest that, in the future, we could train on images annotated with various types of labels—such as boxes, circles, and crosses—as provided by different human annotators. This would enable us to output image labels in the same format specified by prompts.

Generalize to unseen datasets The advantage of in-context learning is that it allows models to adapt to new datasets quickly. MVG achieves 0.84 mIoU on MSD-Liver with only new prompts without any fine-tuning. After fine-tuning MVG with one instance on MSD-Spleen and MSD-Lung, MVG achieves 0.87 mIoU and 0.48 mIoU.

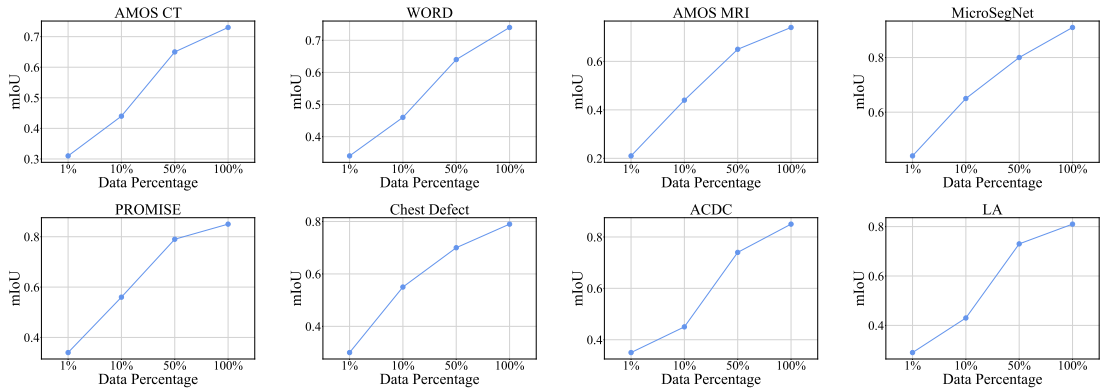

Figure 7: **Impact of training data scale.** We ablate on various scales of the training data (randomly sampled from each dataset), ranging from 1% to 100%.

Method	AMOS CT	WORD	BTCV	AMOS MRI	MicroSegNet	PROMISE	Chest Defect	ACDC	LA
Binary	0.46	0.48	0.48	0.49	0.78	0.78	0.45	0.49	0.52
Pre-defined	0.60	0.62	0.62	0.61	0.89	0.86	0.71	0.74	0.73
Random	0.73	0.74	0.73	0.74	0.91	0.85	0.79	0.85	0.81

Table 4: **Color space for segmentation.** Using semantic masks in a random color space as prompts significantly improves the segmentation performance of our generalist model.

4.3 Ablation Study

Data scalability Nowadays, more and more datasets are available, which motivates us to study whether a scale-up dataset can train a stronger MVG. As shown in Figure 7, we randomly choose 1%, 10%, 50% as comparison with the *full data*. The performance of our MVG consistently improves with the growth of dataset size. These results show the strong dataset scalability of our MVG.

Color space To unify the output space of different segmentation tasks in which the same value from different datasets may have different semantics, we propose to unify the output space with a pre-defined color for each class or random color that we keep the same semantics have the colors. As shown in Table 4, the random color performs much better than the pre-defined color on abdominal segmentation while having a similar performance on prostate segmentation. In particular, our MVG with random color space gains the average result of 0.735 mIoU on abdominal segmentation and improves 0.123 mIoU over that with pre-defined color space. In contrast, our MVG achieves inferior performance with pre-defined or random colors. The random color makes MVGs learn more from the context instead of the color itself and avoid the model being limited by the number of colors.

Isolated and Unified training To validate that our MVG can benefit from large-scale datasets across different tasks. We compare two settings: 1) isolated training: we train different models on different datasets in isolation. Namely, we train 13 models for the 13 datasets. 2) unified training: we train our MVG on all datasets together. Note that both settings have the same model architecture. We report the results in Table 5. The unified model makes significant improvements over the isolated model in all tasks and the improvements reach 0.14 mIoU. Such results indicate that MVG can benefit from large-scale datasets even if this dataset has different annotation semantics which motivates the medical image analysis community to further expand the datasets.

The hybrid training paradigm We ablate the two conditional image generation methods used in our hybrid training paradigm: mask image modeling (MIM) and auto-regressive (AR) training.

While AR training outperforms MIM for image segmentation, this superiority does not extend to other tasks. MIM's suboptimal performance in segmentation, as shown in Table 6, arises from its masking strategy, which

Method	AMOS CT	WORD	BTCV	AMOS MRI	MicroSegNet	PROMISE	Chest Defect	ACDC	LA
Isolated	0.55	0.57	0.57	0.58	0.80	0.77	0.70	0.69	0.68
Unified	0.73	0.74	0.73	0.74	0.91	0.85	0.79	0.85	0.81

Table 5: **Isolated vs. Unified training.** "Isolated" indicates training our MVG individually on each dataset, while "Unified" indicates training on all datasets together.

Method	AMOS CT	WORD	BTCV	AMOS MRI	MicroSegNet	PROMISE	Chest Defect	ACDC	LA
MIM (mask 50%)	0.56	0.48	0.46	0.54	0.70	0.66	0.50	0.50	0.52
MIM (mask 75%)	0.53	0.42	0.44	0.52	0.70	0.63	0.48	0.50	0.51
Auto-regressive	0.73	0.74	0.73	0.74	0.91	0.85	0.79	0.85	0.81

Table 6: **Auto-regressive training boosts in-context segmentation.** Randomly masking can harm in-context segmentation task, especially when it results in the complete removal of small organs. Auto-regressive training addresses this weakness and makes much better performance than MIM.

can disrupt the preservation of global contextual information crucial for delineating anatomical structures, such as the spatial relationships among abdominal organs. This aligns with findings from (Xie et al., 2023b), which suggest that MIM is better suited for capturing local details but struggles with maintaining global context. However, for tasks like inpainting and denoising, where refining local details takes precedence over preserving global context, MIM consistently outperformed AR training as shown in Table 7.

Based on these insights and the results in Table 7, we adopted a hybrid training strategy: For segmentation tasks, we use AR training exclusively. For all other tasks, we allocate 90% of training iterations to MIM and 10% to AR training.

Based on these insights, we adopted a hybrid training strategy to harness the strengths of both methods. For segmentation tasks, we rely exclusively on AR training, as it effectively preserves the global spatial and contextual information essential for accurate delineation in medical images. For other tasks, such as inpainting and denoising, we allocate 90% of the training iterations to MIM and 10% to AR training. This approach ensures that the inherent spatial and contextual information of medical images is utilized to enhance in-context learning and segmentation performance while maintaining the ability to refine local details.

4.4 Discussion and Limitation.

In this research work, our primary focus is on demonstrating how our pipeline enables the integration of multiple medical vision tasks into a unified image-to-image generation framework, representing a novel vision-centric approach to medical generalist models. While generalist models currently underperform specialized models across various benchmarks—spanning both natural and medical imaging domains—they offer transformative potential by unifying multiple modalities, tasks, and datasets into a single framework. This new paradigm provides far greater flexibility compared to task-specific models—unlike specialized models which are limited to segmentation tasks, MVG can be directly trained across a wide range of tasks and easily adapted to new datasets with minimal task-specific data.

In addition, we offer an in-depth analysis of the role of mask image modeling and auto-regressive training in developing a generalist medical model—an investigation that has not been previously conducted. We also ablate the role of objective function, data/task balancing, and analyze the new emerging generalization to new discriminative tasks such as detection, aiming to offer practical guidelines for training future generalist models in the medical imaging domain. Furthermore, we highlight the utility of MVG, particularly in comparison to prior frameworks like UniverSeg and MedSAM, showcasing its enhanced flexibility and adaptability to customized user datasets.

Generalization to new medical imaging modalities Currently, MVG struggles to handle out-of-domain data, such as pathology images, which differ significantly from the radiology modalities (e.g., CT and MRI) used in its training dataset. This limitation arises from the domain and modality gaps, making MVG

Method	Cross-modal synthesis	Inpainting	Denoise
MIM	0.019	0.006	0.018
Autoregressive	0.020	0.006	0.019

Table 7: **Mask Image Modeling (MIM) vs. Autoregressive on inpainting and denoising.** MIM performs better if the task like inpainting and denoising reqiures more local details over global context.

unsuitable for direct application to such datasets without additional pretraining on these data types. To address this issue, future work will focus on enhancing the diversity of imaging modalities included during the pretraining stage by incorporating pathology images and other underrepresented medical imaging data. Expanding the variety of training data is expected to improve MVG's generalization capabilities, enabling it to perform robustly across a wider range of medical imaging datasets and enhancing its applicability in diverse clinical scenarios.

The 2D framework Our current framework operates on 2D images, similar to approaches like UniverSeg (Butoi et al., 2023). While we acknowledge the critical importance of 3D contextual information for accurately analyzing anatomical structures and the limitations of 2D analysis in capturing volumetric relationships, transitioning directly to 3D models presents significant computational challenges that require careful, non-trivial design. Exploring the integration of 2.5D or fully 3D models into MVG represents an important direction for future research.

Generalist vs. Specialist Models Generalist models typically require more computational resources because they need to handle a wide range of tasks simultaneously. Nevertheless, generalist models still have non-negligible advantages over specialized models. They provide versatility by handling multiple tasks or domains within a single framework, reducing the need for separate models and minimizing development time. Their ability to leverage shared knowledge across tasks improves data efficiency and facilitates scalability as new tasks emerge. Additionally, generalist models often excel in transfer learning, offering cost-effective solutions by integrating various functionalities and insights into one unified system.

Performance-wise, although generalist models currently underperform relative to specialized models across various benchmarks, including both natural and medical images, the robust scalability suggests that, with increased data and computational resources, our generalist approach may have the potential to surpass specialized models in the future

Our pipeline unlocks the possibility of integrating multiple medical vision tasks into a unified image-to-image generation framework. While generalist models may not yet outperform domain-specific models like U-Net or nnUNet in segmentation, they offer greater flexibility. MVG, unlike specialist models, can be directly trained across diverse tasks and adapted to new datasets with minimal task-specific samples, whereas models like U-Net and nnUNet are limited to segmentation and cannot handle other tasks.

5 Conclusion

In this work, we present MVG, a versatile model capable of handling various medical imaging tasks, including cross-modal synthesis, segmentation, denoising, and inpainting, within a unified image-to-image generation framework. MVG employs an in-context generation strategy to standardize inputs and outputs as images, allowing flexible task unification across various modalities and datasets. A hybrid approach combining masked image modeling and autoregressive training proved the most effective. To thoroughly assess MVG's potential and limitations, we also curate the first comprehensive generalist medical vision benchmark consisting of 13 datasets across 4 imaging modalities, including CT, MRI, X-ray, and Micro-ultrasound. Experiment results demonstrate that MVG consistently outperforms existing vision generalists. Benefiting from the in-context learning scheme, MVG demonstrates exceptional flexibility, scalability, and potential for generalization to unseen datasets with minimal samples. We will make our code and benchmark publicly available to encourage future research in medical AI generalists. We believe MVG can serve as a stepstone to make medical imaging tools more accessible to clinical researchers, other scientists, or beyond, by lowering the bar of machine learning expertise, computational resources, and human labor. We discuss the limitation in appendix.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

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.

Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954–961, 2019.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023.

Olivier Bernard, Alain Lalande, Clément Zotti, Frédéric Cervenansky, Xin Yang, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37:2514–2525, 2018.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020a.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *preprint arXiv:2005.14165*, 2020b.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *ICCV*, 2023.

Sema Candemir and Sameer Antani. A review on lung boundary detection in chest x-rays. *International journal of computer assisted radiology and surgery*, 14:563–576, 2019.

Chen Chen, Wenjia Bai, and Daniel Rueckert. Multi-task learning for left atrial segmentation on ge-mri. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9*, pp. 292–301. Springer, 2019.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021a.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020.

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021b.

Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022.

Chi-Tung Cheng, Jinzheng Cai, Wei Teng, Youjing Zheng, Yu-Ting Huang, Yu-Chao Wang, Chien-Wei Peng, Youbao Tang, Wei-Chen Lee, Ta-Sen Yeh, et al. A flexible three-dimensional heterophase computed tomography hepatocellular carcinoma detection algorithm for generalizable and practical screening. *Hepatology Communications*, 2022a.

Junlong Cheng, Shengwei Tian, Long Yu, Chengrui Gao, Xiaojing Kang, Xiang Ma, Weidong Wu, Shijia Liu, and Hongchun Lu. Resganet: Residual group attention network for medical image classification and segmentation. *Medical Image Analysis*, 76:102313, 2022b.

Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. Deep learning based synthesis of mri, ct and pet: Review and analysis. *Medical Image Analysis*, pp. 103046, 2023.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *preprint arXiv:2010.11929*, 2020.

Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. A review of deep learning based methods for medical image multi-organ segmentation. *Physica Medica*, 85:107–122, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021.

Tianyu Hua, Yonglong Tian, Sucheng Ren, Michalis Raptis, Hang Zhao, and Leonid Sigal. Self-supervision through random segments with autoregressive coding (randsac). In *The Eleventh International Conference on Learning Representations*, 2022.

Yuankai Huo, Jinzheng Cai, Chi-Tung Cheng, Ashwin Raju, Ke Yan, Bennett A Landman, Jing Xiao, Le Lu, Chien-Hung Liao, and Adam P Harrison. Harvesting, detecting, and characterizing liver lesions from large-scale multi-phase CT data via deep dynamic texture learning. *arXiv preprint arXiv:2006.15691*, 2020.

Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219, 2015.

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.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *NeurIPS*, 2022.

Hongxu Jiang, Muhammad Imran, Preethika Muralidharan, Anjali Patel, Jake Pensa, Muxuan Liang, Tarik Benidir, Joseph R Grajo, Jason P Joseph, Russell Terry, et al. Microsegnet: a deep learning approach for prostate segmentation on micro-ultrasound images. *Computerized Medical Imaging and Graphics*, 112: 102326, 2024.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *ArXiv*, 2023.

Roger Y Kim, Jason L Oke, Lyndsey C Pickup, Reginald F Munden, Travis L Dotson, Christina R Bellinger, Avi Cohen, Michael J Simoff, Pierre P Massion, Claire Filippini, et al. Artificial intelligence tool for assessment of indeterminate pulmonary nodules detected with CT. *Radiology*, pp. 212182, 2022.

Florian Kofler, Felix Meissen, Felix Steinbauer, Robert Graf, Eva Oswald, Ezequiel de da Rosa, Hongwei Bran Li, Ujjwal Baid, Florian Hoelzl, Oezguen Turgut, et al. The brain tumor segmentation (brats) challenge 2023: Local synthesis of healthy brain tissue via inpainting. *arXiv preprint arXiv:2305.08992*, 2023.

Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, pp. 12, 2015.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *arXiv preprint arXiv:2210.14215*, 2022.

Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *NeurIPS*, 2024.

Hongwei Bran Li, Gian Marco Conte, Syed Muhammad Anwar, Florian Kofler, Ivan Ezhov, Koen van Leemput, Marie Piraud, Maria Diaz, Byrone Cole, Evan Calabrese, et al. The brain tumor segmentation (brats) challenge 2023: Brain mr image synthesis for tumor segmentation (brasyn). *ArXiv*, 2023.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *arXiv preprint arXiv:2111.02403*, 2021.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

C McCollough, B Chen, D Holmes, X Duan, Z Yu, L Xu, S Leng, and J Fletcher. Low dose ct image and projection data (ldct-and-projection-data)(version 4). *Med. Phys*, 48:902–911, 2021.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL `https://api.semanticscholar.org/CorpusID:257532815`.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023b.

Guoyang Xie, Yawen Huang, Jinbao Wang, Jiayi Lyu, Feng Zheng, Yefeng Zheng, and Yaochu Jin. Cross-modality neuroimage synthesis: A survey. *ACM computing surveys*, 56(3):1–28, 2023a.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14475–14485, 2023b.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3): 036501–036501, 2018.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023.

Tianyi Zhao, Kai Cao, Jiawen Yao, Isabella Nogues, Le Lu, Lingyun Huang, Jing Xiao, Zhaozheng Yin, and Ling Zhang. 3d graph anatomy geometry-integrated network for pancreatic mass segmentation, diagnosis, and quantitative patient management. In *CVPR*, 2021.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompt. *arXiv preprint arXiv:2312.17183*, 2023.

Lingting Zhu, Zeyue Xue, Zhenchao Jin, Xian Liu, Jingzhen He, Ziwei Liu, and Lequan Yu. Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In *MICCAI*, 2023.

Zhuotun Zhu, Yingda Xia, Lingxi Xie, Elliot K Fishman, and Alan L Yuille. Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. In *MICCAI*, 2019.