This supplementary material is organized as follows: Sec. A provides more details about the datasets utilized in model training. More implantation details about three networks and downstream tumor segmentation tasks are provided in Sec. B. Sec. C contains additional visualizations of synthetic data and ablation studies.

A. Dataset Details

A.1. MAISI VAE

For the foundational 3D VAE in MAISI, we include a diverse dataset comprising 37,243 CT volumes for training and 1,963 CT volumes for validation, covering the chest, abdomen, and head and neck regions. Additionally, we include 17,887 MRI volumes for training and 940 MRI volumes for validation, spanning the brain, skull-stripped brain, chest, and below-abdomen regions. The training data were sourced from various repositories, including TCIA COVID-19 Chest CT, TCIA Colon Abdomen CT, MSD03 Liver Abdomen CT, LIDC Chest CT, TCIA Stony Brook COVID Chest CT, NLST Chest CT, TCIA Upenn GBM Brain MR, AOMIC Brain MR, QTIM Brain MR, TCIA Acrin Chest MR, and TCIA Prostate MR. This extensive and varied dataset not only ensures that our model is exposed to a broad range of anatomical regions but also supports its application to both MRI and CT images.

The details of MAISI VAE training data are shown in Table S1.

Dataset Name	**Number of Training Data**	**Number of Validation Data**
Covid 19 Chest CT	722	49
TCIA Colon Abdomen CT	1522	77
MSD03 Liver Abdomen CT	104	0
LIDC chest CT	450	24
TCIA Stony Brook Covid Chest CT	2644	139
NLST Chest CT	31801	1674
TCIA Upenn GBM Brain MR (skull-stripped)	2550	134
Aomic Brain MR	2630	138
QTIM Brain MR	1275	67
Acrin Chest MR	6599	347
TCIA Prostate MR Below-Abdomen MR	928	49
Aomic Brain MR, skull-stripped	2630	138
QTIM Brain MR, skull-stripped	1275	67
Total CT	37243	1963
Total MRI	17887	940

Table S1. MAISI VAE Dataset Information

A.2. MAISI Diffusion

The datasets for developing the Diffusion model used in MAISI comprise 10,277 CT volumes from 24 distinct datasets, encompassing various body regions and disease patterns. Table S2 provides a summary of the number of volumes for each dataset. For compatibility with the shape requirement of the U-shape network, we resample the dimensions of volumes to multiples of 128. Fig. S1 visualizes the characteristics and spatial complexity of the data involved in training the diffusion model.

A.3. MAISI ControlNet

The ControlNet training dataset for **MAISI CT Generation** discussed in Sec. 4.4 contains 6,330 CT volumes (5,058 and 1,272 volumes are used for training and validation, respectively) across 20 datasets and covers different body regions and diseases. Table S3 summarizes the number of volumes for each dataset.

Dataset name	**Number of volumes**
AbdomenCT-1K	789
AeroPath	15
AMOS22	240
autoPET23 (testing only)	200
Bone-Lesion	223
BTCV	48
COVID-19	524
CRLM-CT	158
CT-ORG	94
CTPelvic1K-CLINIC	94
LIDC	422
MSD Task03	88
MSD Task06	50
MSD Task07	224
MSD Task08	235
MSD Task09	33
MSD Task10	87
Multi-organ-Abdominal-CT	65
NLST	3109
Pancreas-CT	51
StonyBrook-CT	1258
TCIA_Colon	1437
TotalSegmentatorV2	654
VerSe	179
Total	10277

Table S2. MAISI DM Dataset Information

Dataset name	**Number of volumes**
AbdomenCT-1K	789
AeroPath	15
AMOS22	240
Bone-Lesion	237
BTCV	48
CT-ORG	94
CTPelvic1K-CLINIC	94
LIDC	422
MSD Task03	105
MSD Task06	50
MSD Task07	225
MSD Task08	235
MSD Task09	33
MSD Task10	101
Multi-organ-Abdominal-CT	64
Pancreas-CT	51
StonyBrook-CT	1258
TCIA_Colon	1436
TotalSegmentatorV2	654
VerSe	179
Total	6330

Table S3. MAISI ControlNet Dataset Information

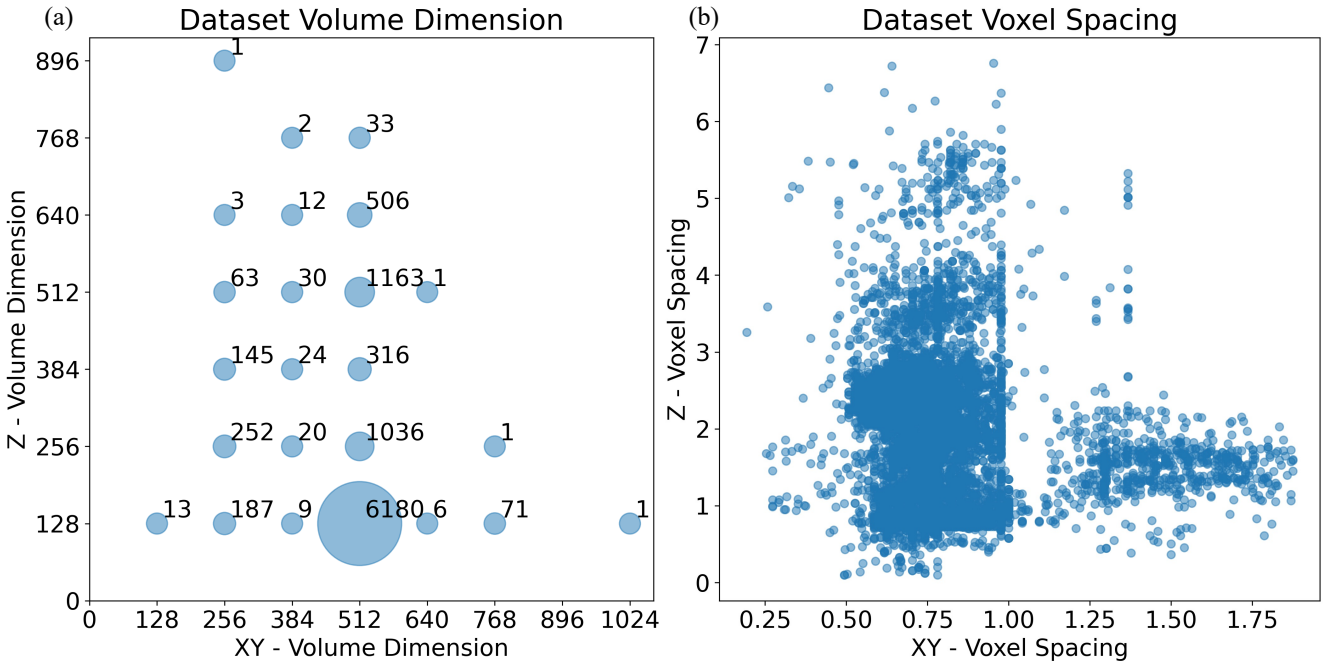

Figure S1. The characteristics of the datasets utilized for the MAISI Diffusion Model are detailed through two subplots. Subplot (a) illustrates the volume dimensions of the datasets, providing insight into the variability and range of sizes used in the training data. Subplot (b) presents the voxel spacing in millimeters for each data point, emphasizing the spatial configuration within the CT scans. Notably, in CT imaging, the X and Y directions typically share identical dimensions and spacing, so they are represented on a single axis in both subplots.

B. Additional Implementation Details

MAISI VAE. To establish the VAE as a foundational model, we employ an extensive range of data augmentation techniques. For CT images, intensities are clipped to a Hounsfield Unit (HU) range of -1000 to 1000 and normalized to a range of [0,1]. For MR images, intensities were normalized such that the 0th to 99.5th percentile values were scaled to the range [0,1]. For MR images, we applied intensity augmentations including random bias field, random Gibbs noise, random contrast adjustment, and random histogram shifts. Both CT and MR images underwent spatial augmentations, such as random flipping, random rotation, random intensity scaling, random intensity shifting, and random upsampling or downsampling.

The MAISI VAE model is trained with 8 32G V100 GPU. It is initially trained for 100 epochs using small, randomly cropped patches of size [64,64,64]. This approach is adopted to improve the model's ability to generalize to images with partial volume effects. After this initial phase, training is continued for an additional 200 epochs using larger patches of size [128,128,128], which allows the model to capture more contextual information and improve overall accuracy.

The MAISI VAE is used to compress the latent features that will be employed in latent diffusion models, where having a well-structured and meaningful latent space is crucial for effective diffusion dynamics. Therefore, during MAISI VAE training, we adjust the weight of the KL loss to ensure the standard deviation remains between 0.9 to 1.1. This calibration balances the model's focus between accurate data reconstruction and adherence to the prior distribution. As the MAISI VAE is intended to serve as a foundational model, maintaining this balance also helps to prevent over-fitting [28].

MAISI Diffusion. Data preprocessing for diffusion model training involves applying a series of precise transformations to the image data, including loading the images, ensuring the correct channel structure, adjusting the orientation according to the "RAS" axcode, and scaling intensity values from -1000 to 1000 to normalize the data between 0 and 1. The process further refines the images by adjusting dimensions to the nearest multiple of 128, recording the new spatial details, using trilinear interpolation. Then each image is passed through a pre-trained autoencoder, generating a compressed latent representation that is saved for subsequent model training. The diffusion model requires additional input attributes, including output dimensions, output spacing, and top/bottom body region indicators. These dimensions and spacing are extracted from the header information of the training images. The top and bottom body regions can be identified either through manual inspection or by using segmentation tools such as TotalSegmentator [63] and VISTA3D [26]. These regions are encoded as 4-dimensional one-hot vectors: the head and neck region is represented by $[1, 0, 0, 0]$, the chest by $[0, 1, 0, 0]$, the abdomen by $[0, 0, 1, 0]$, and the lower body (below the abdomen) by $[0, 0, 0, 1]$. These additional input attributes are stored in a separate configuration file. In this example, it is assumed that the images encompass the chest and abdomen regions.

Next, the diffusion model training process begins with an initial learning rate of $1e^{-4}$, a batch size of 1, and spans 200 epochs. To ensure the data is optimally prepared for training, various transformations are applied to the image inputs. The U-Net architecture is employed for noise prediction, with distributed computing utilized to enhance efficiency when multiple GPUs are available. The Adam optimizer is responsible for adjusting the model's parameters, while a polynomial learning rate scheduler controls the update rate over training steps. Noise is systematically introduced to the input data by the noise scheduler, and the model iteratively refines its predictions using an L1 loss function to minimize this noise. Mixed precision training and gradient scaling are implemented to optimize memory usage and computational performance.

MAISI ControlNet. We train a versatile ControlNet Model (MAISI CT Generation task in Sec. 4.4) to support all five types of tumors using the datasets summarized in Table S3. The data preprocessing protocol is the same in the training of the MAISI Diffusion Model. The Adam optimizer is employed for training purposes, with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set at 0.0001, with the polynomial learning rate decay. The batch size is set to 1 per GPU. Training is performed on a server with 8 A100 GPUs with about 10k optimization steps. For the MAISI Inpainting task, we employ the same hyperparameters for training but only use datasets with supported tumor types, including MSD Task03 [3] (liver tumor), Task06 [3] (lung tumor), Task07 [3] (pancreas tumor).

Downstream tumor segmentation. The implementation of all tumor segmentation models is based on the Auto3DSeg[4] pipeline. Auto3DSeg is an auto-configuration pipeline designed for 3D medical image segmentation, utilizing MONAI [6]. The pipeline begins with data analysis to extract global information from the dataset, followed by algorithm generation based on data statistics and predefined templates. It then proceeds to model training to obtain optimal checkpoints. All used tumor dataset is split into 80% for training and 20% for testing. The training set is further divided into five folds for 5-fold cross-validation. We report the segmentation performance on the holdout testing set. For the MAISI CT Generation task, we generate synthetic data from augmented real masks containing tumors. Fig. S2 shows an example of mask augmentation for a case with the lung tumor. For the MAISI Inpainting task, we follow the same setting in DiffTumor [10] and use the

[4]https://monai.io/apps/auto3dseg

provided healthy cases in the open-source repository[5] to generate synthetic data with tumors. For both tasks, the amount of synthesized data is equivalent to the original dataset size for each tumor type. We explore the impact of using different amounts of synthetic data for data augmentation in Supplementary Sec. C.

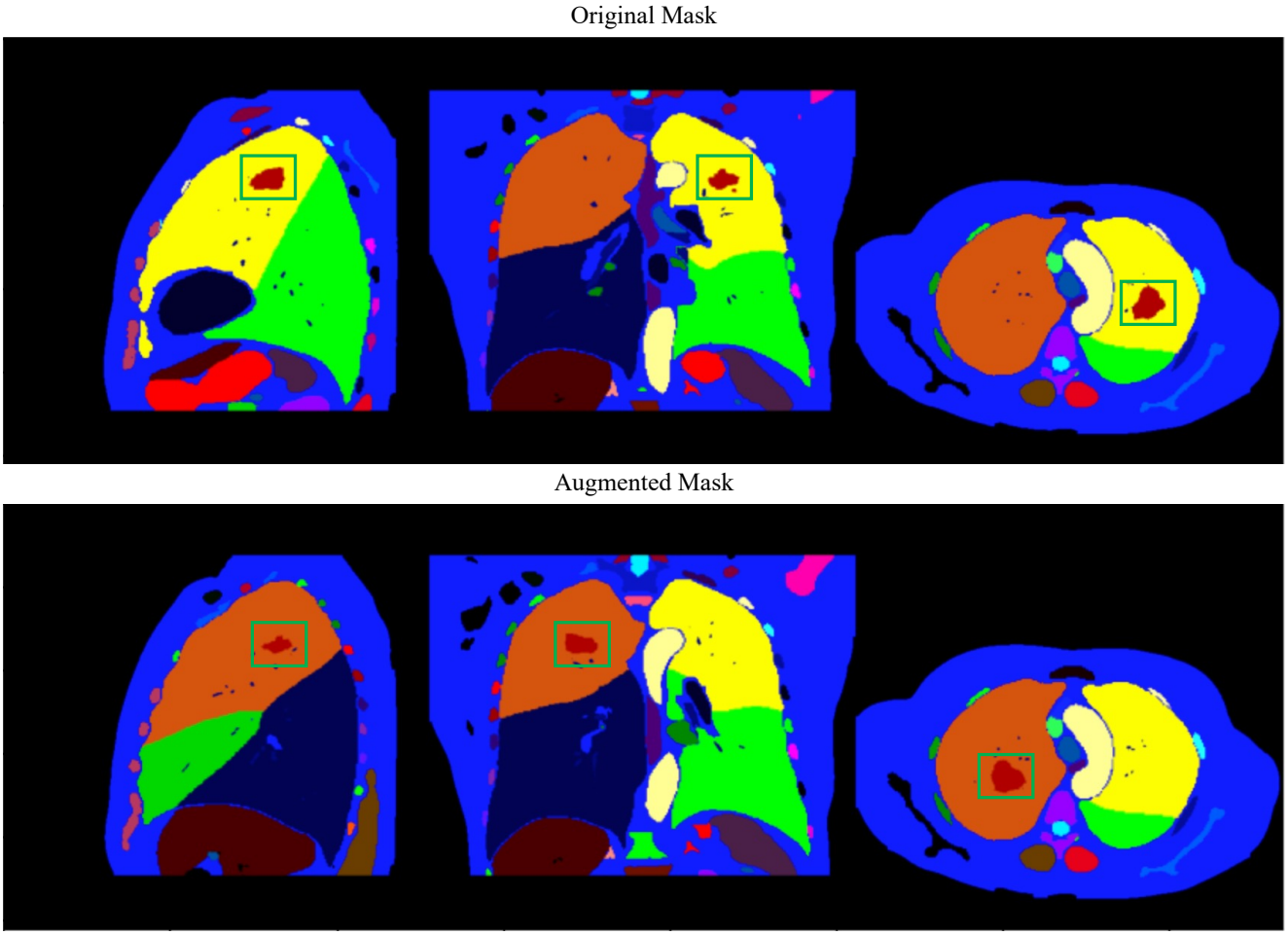

Figure S2. The example lung tumor mask and corresponding augmented mask. The green boxes highlight the tumor regions in different views.

[5] https://github.com/MrGiovanni/DiffTumor

C. Supplementary Experiment Results

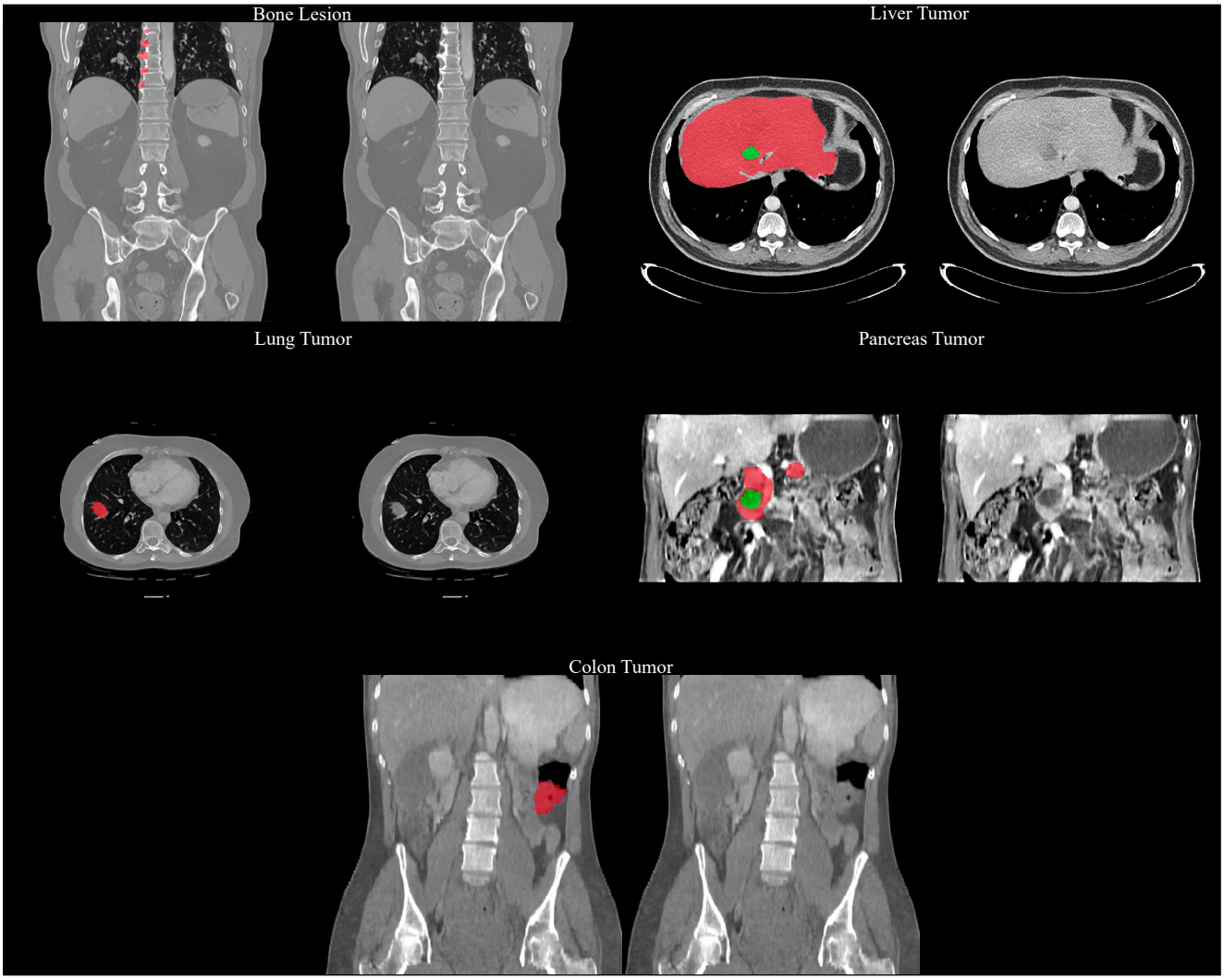

Figure S3. The example of generated images from MAISI CT Generation task.

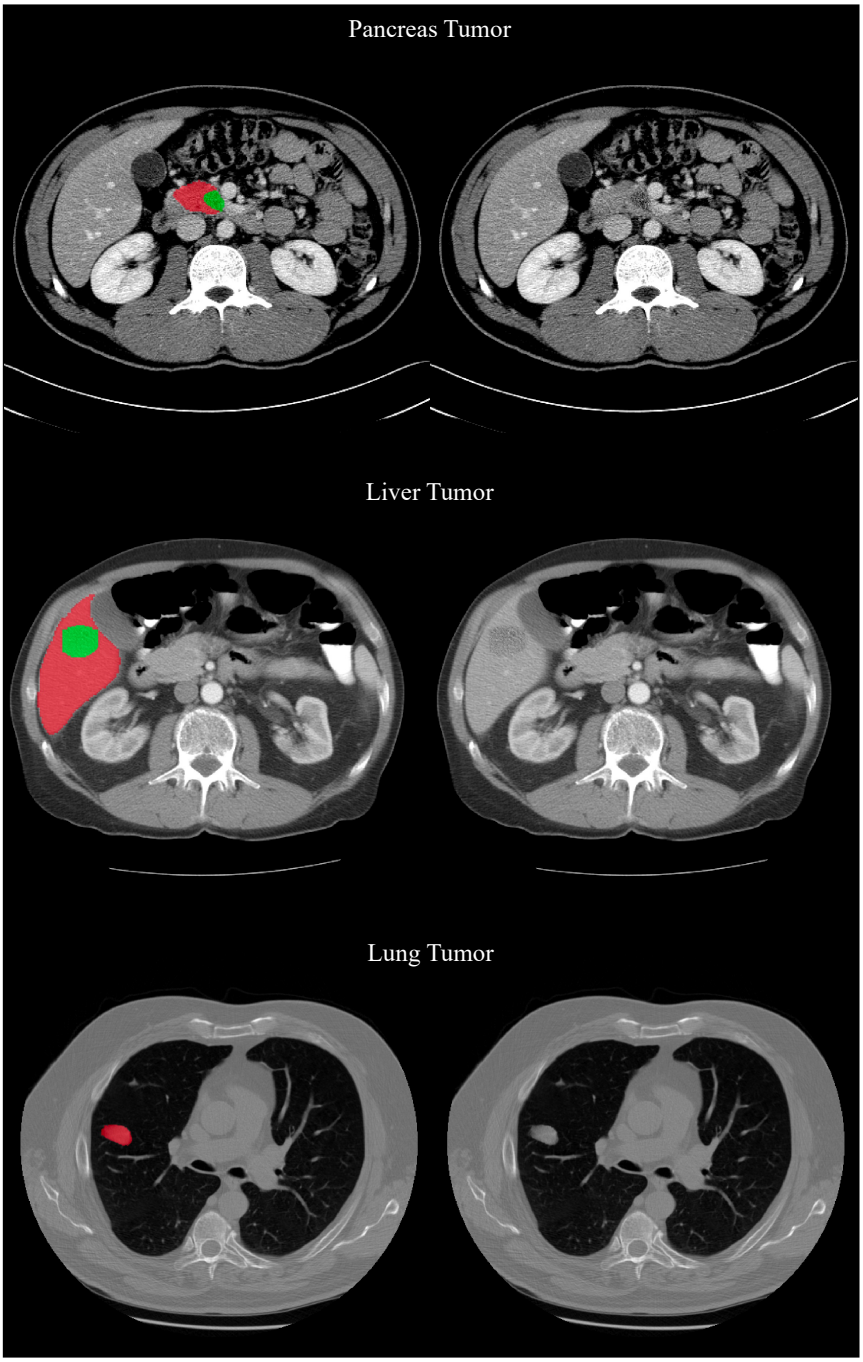

Figure S4. The example of generated images from MAISI Inpainting task.

MSD Task06	Real v.s. Synthetic	fold 0	fold 1	fold 2	fold 3	fold 4	Avg.	Improvement
Real Only	1:0	0.494	0.601	0.535	0.674	0.599	0.581	-
MAISI CT Generation	1:1	0.585	0.649	0.631	0.647	0.664	0.635	5.5%
MAISI CT Generation	1:0.5	0.640	0.593	0.606	0.639	0.644	0.624	4.4%
MAISI CT Generation	1:1.5	0.641	0.658	0.586	0.645	0.666	0.639	5.8%
MSD Task07	Real v.s. Synthetic	fold 0	fold 1	fold 2	fold 3	fold 4	Avg.	Improvement
Real Only	1:0	0.423	0.463	0.414	0.42	0.444	0.433	-
MAISI CT Generation	1:1	0.504	0.448	0.467	0.482	0.508	0.482	4.9%
MAISI CT Generation	1:0.5	0.465	0.463	0.423	0.447	0.478	0.455	2.2%
MAISI CT Generation	1:1.5	0.466	0.481	0.465	0.480	0.467	0.471	3.9%

Table S4. The ablation study examines the effect of varying amounts of synthetic data in data augmentation experiments. The 'Improvement' column reports the percentage of relative improvement compared to experiments using only real data. We conduct this ablation study on the smallest dataset (MSD Task06) and the largest dataset (MSD Task07) across five tumor types. Our empirical results suggest that using a synthetic dataset equivalent in size to the original dataset is an effective choice for data augmentation.

	Liver	Spleen	Left Kidney	Right Kidney	Stomach	Gallbladder	Esophagus	Pancreas	Duodenum	Colon	Small Bowel	Bladder
Real Data	0.95	0.94	0.93	0.93	0.90	0.75	0.76	0.80	0.69	0.76	0.80	0.91
Synthetic Data	0.93	0.93	0.95	0.95	0.88	0.47	0.73	0.70	0.54	0.73	0.74	0.86

Table S5. Segmentation performance on synthetic data. Synthetic data is generated using the MAISI CT Generation task and evaluated with the VISTA 3D [26] segmentation model. DSC are presented for both synthetic and real data on the unseen WORD [41] dataset. The results demonstrate that the segmentation model achieves comparable performance on major organs (*e.g.*, liver, spleen, kidney) for both synthetic and real data. However, smaller organs (*e.g.*, gallbladder, duodenum, pancreas) show a more pronounced performance gap between synthetic and real data. Addressing this gap presents a promising direction for future research.