MRGen: Segmentation Data Engine for Underrepresented MRI Modalities

Appendix

Contents

A. Preliminaries on Diffusion Models

Diffusion Models [9] are a class of deep generative models that convert Gaussian noise into structured data samples through an iterative denoising process. These models typically comprise a forward diffusion process and a reverse denoising process.

Specifically, the forward diffusion process progressively introduces Gaussian noise into an image ($\mathbf{x}_0$) via a Markov process over T steps. Let $\mathbf{x}_t$ represent the noisy image at step t. The transition from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$ can be formulated as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

Here, $\beta_t \in (0,1)$ represents pre-determined hyperparameters that control the variance at each step. By defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, the properties of Gaussian distributions and the reparameterization trick allow for a refined expression:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

This insight provides a concise expression for the forward process with Gaussian noise $\boldsymbol{\epsilon}$ as: $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$.

Diffusion models also encompass a reverse denoising process that reconstructs images from noise. A UNet-based model [25] is typically utilized to learn the reverse diffusion process p_θ, represented as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$$

Here, μ_θ represents the predicted mean of the Gaussian distribution, derived from the estimated noise $\boldsymbol{\epsilon}_\theta$ as:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t))$$

Building on this foundation, **Latent Diffusion Models** [24] adopt a Variational Autoencoder (VAE [16]) to project images into a learned, compressed, low-dimensional latent space. The forward diffusion and reverse denoising processes are then performed on the latent codes ($\mathbf{z}$) within this latent space, significantly reducing computational cost and improving efficiency.

B. Details of MRGen-DB & Synthetic Data

This section provides additional details about our curated **MRGen-DB** dataset. In Sec. B.1, we elaborate on the implementation details of the automatic annotation pipeline; and in Sec. B.2, we present more comprehensive data statistics. Moreover, in Sec. B.3, we provide statistics on the MRGen-synthesized data used for downstream segmentation models training.

B.1. Automatic Annotations

We employ an automated annotation pipeline to annotate our MRGen-DB dataset, ensuring that the templated text prompts contain sufficient and clinically relevant information to distinguish distinct modalities, regions, and organs. This process primarily consists of two precise and controllable components: human body region classification and modality explanation, which will be detailed as follows.

Region classification. Considering the wide range and variability of abdominal imaging, we adopt the off-the-shelf BiomedCLIP [31] image encoder to encode all 2D slices, and the BiomedCLIP text encoder to encode predefined text descriptions of six abdominal regions. Based on the cosine similarity between the image and text embeddings, the 2D slices are classified into one of the six categories, including *Upper Thoracic Region*, *Middle Thoracic Region*, *Lower Thoracic Region*, *Upper Abdominal Region*, *Lower Abdominal Region*, and *Pelvic Region*. For text encoding, we use a templated text prompt as input:

> *This is a radiology image that shows $region$ of a human body, and probably contains $organ$.*

Here, $region$ and $organ$ represent the items in the following list:

> *(region, organ) = [('Upper Thoracic Region', 'lung, ribs and clavicles'), ('Middle Thoracic Region', 'lung, ribs and heart'), ('Lower Thoracic Region', 'lung, ribs and liver'), ('Upper Abdominal Region', 'liver, spleen, pancreas, kidney and stomach'), ('Lower Abdominal Region', 'kidney, small intestine, colon, cecum and appendix'), ('Pelvic Region', 'rectum, bladder, prostate/uterus and pelvic bones')]*

Modality explanation. To capture the correlations and distinctions among various modality labels, we leverage GPT-4 [1] to generate free-text descriptions detailing the signal intensities of *fat*, *muscle*, and *water* for each modality label. This helps the model better understand the imaging characteristics of distinct modalities. The prompt we use is as follows:

As a senior doctor and medical imaging researcher, please help me map radiological imaging modalities to the signal intensities of fat, muscle, and water, as well as their corresponding brightness levels. Please provide the answer in the following format: fat {} signal, muscle {} signal, water {} signal, fat {}, muscle {}, water {}. Now, tell me the attributes of $modality$.

To ensure reliability and accuracy, we have randomly and uniformly sampled approximately 2% (5K out of 250K) of region annotations and 20% (60 out of ∼300) of modality attribute annotations for manual verification, achieving high accuracies of 95.33% and 91.67%. Furthermore, the effectiveness in downstream tasks also validates the quality of automatic annotations.

B.2. Dataset Statistics

In this section, we present more detailed statistics about our curated MRGen-DB dataset, including the unannotated image-text pairs from *Radiopaedia*[1], as well as the mask-annotated data sourced from various open-source datasets.

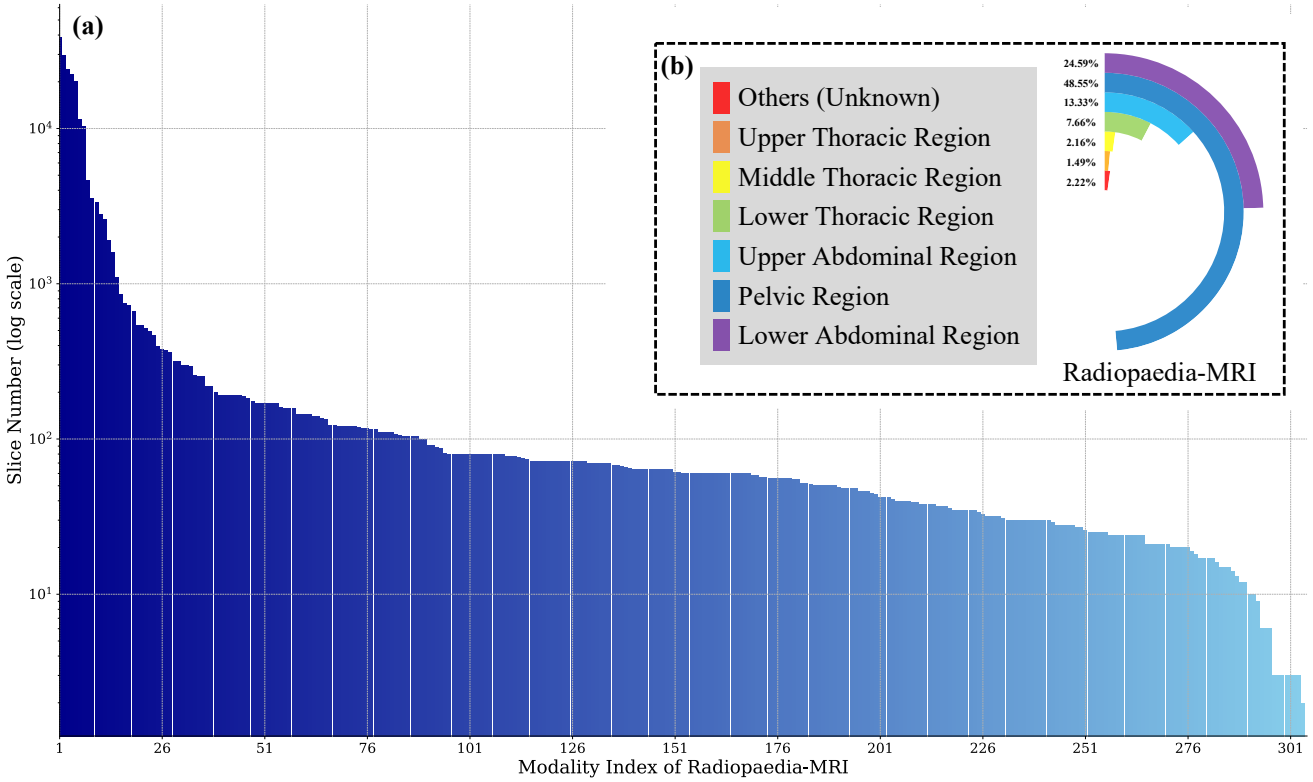


Figure 1. **Data Statistics of *Radiopaedia-MRI*.** (a) Distribution of slice counts across various modalities in *Radiopaedia-MRI*; (b) Proportional distribution of slices across different regions in *Radiopaedia-MRI*.

Data without mask annotations. For the image-text pairs from *Radiopaedia-MRI*, which are used for training the autoencoder and text-guided generation, we allocate 1% of the data as a test set to evaluate reconstruction and generation performance, maximizing the amount of data available for pretraining. As a result, 202,988 samples are used for training, and the test set consists of 2,051 samples. We conduct a statistical analysis of the distribution of modalities in *Radiopaedia-MRI*, as presented in Figure 1 (a). The free-text modality labels cover approximately 300 categories, providing a diverse set of MRI modalities that form a crucial foundation for MRGen to learn text-guided generation and expand its mask-conditioned generation capabilities towards modalities originally lacking mask annotations. Furthermore, the distribution of images across different regions in *Radiopaedia-MRI* is presented in Figure 1 (b).

Data with mask annotations. Following the SAT [33], we split the data with mask annotations into training and test sets, as detailed in Table 1. For dataset pairs comprising different datasets, we use their shared organs as the segmentation targets.

[1]radiopaeida.org

Dataset	Organs	Modality	Train			Test		
			# Vol.	# Slc.	# Slc. w/ mask	# Vol.	# Slc.	# Slc. w/ mask
PanSeg [32]	Pancreas	T1-weighted	309	14,656	5,961	75	3,428	1,400
		T2-weighted	305	12,294	5,106	77	2,982	1,312
MSD-Prostate [3]	Prostate	T2-weighted	26	492	100	6	110	83
		ADC	26	492	100	6	110	83
CHAOS-MRI [13]	Liver, Right Kidney,	T1-weighted	32	1,018	770	8	276	230
	Left Kidney, Spleen	T2-SPIR	16	503	388	4	120	104
PROMISE12 [17]	Prostate	T2-weighed	40	1,137	645	10	240	133
LiQA [18]	Liver	T1-weighted	24	1,718	1,148	6	467	298
Total	/	/	**778**	**36,710**	**14,218**	**192**	**7,733**	**3,643**

Table 1. **Details of Segmentation-annotated Datasets in MRGen-DB.** Here, # Vol. represents the number of 3D Volumes, # Slc. denotes the number of 2D slices, and # Slc. w/ mask indicates the number of 2D slices with mask annotations.

B.3. Synthetic Data Statistics

This section presents the statistics of target-domain training samples synthesized by MRGen across various experimental settings, as presented in Table 2. Concretely, we use mask annotations from the entire source-domain dataset (including both training and test sets) as input conditions to generate target-domain images, forming image-mask training pairs. Exceptions include: (i) for the MSD-Prostate [3] dataset, where images of T2 and ADC modalities have already been registered, we restrict inputs to the source-domain training set to prevent data leakage; and (ii) for dataset pairs with CHAOS-MRI-T1 [13] as the target domain, each source-domain mask is used twice to synthesize both T1-InPhase and T1-OutofPhase data. This setup is consistent across all baselines. Additionally, with our proposed autofilter pipeline, MRGen generates 20 candidate images per mask and selects the top two that meet the predefined thresholds. If no samples satisfy the thresholds, all thresholds will be relaxed by 0.10, and the sample of the highest quality is chosen, ensuring full exploitation of source-domain masks. Otherwise, all low-quality generated samples are discarded to avoid noisy data.

Source Dataset	Source Modality	Target Dataset	Target Modality	# Slices ($\mathcal{D}_s$)	# Synthetic Data
CHAOS-MRI	T1	CHAOS-MRI	T2-SPIR	1,294	433
CHAOS-MRI	T2-SPIR	CHAOS-MRI	T1	607	1,118
MSD-Prostate	T2	MSD-Prostate	ADC	492	775
MSD-Prostate	ADC	MSD-Prostate	T2	492	745
PanSeg	T1	PanSeg	T2	18,084	2,160
PanSeg	T2	PanSeg	T1	15,276	2,215
LiQA	T1	CHAOS-MRI	T2-SPIR	2,185	2,267
CHAOS-MRI	T2-SPIR	LiQA	T1	607	636
MSD-Prostate	ADC	PROMISE12	T2	602	742
PROMISE12	T2	MSD-Prostate	ADC	1,377	1,077

Table 2. **Synthetic Data Statistics.** Here, # Slices ($\mathcal{D}_s$) denotes the number of source-domain samples under each experimental setting, which serves as input for translation-based baselines. Moreover, # Synthetic Data represents the total volume of data generated by MRGen.

C. Implementation Details

In this section, we will provide a comprehensive explanation of the implementation details discussed in the paper. Concretely, Sec. C.1 describes the preprocessing and augmentation strategies applied to the training data. Sec. C.2 elaborates on the details of the autofilter pipeline. Finally, Sec. C.3 outlines the implementation details of various baselines.

C.1. Preprocessing & Augmentation

Data preprocessing. To ensure consistency across data from various sources and modalities, we apply tailored preprocessing strategies as follows: (i) For data from *Radiopaedia-MRI*, the images are directly rescaled to the range [0, 1]; (ii) For MR im-

ages with mask annotations, intensities are clipped to the 0.5 and 99.5 percentiles and rescaled to [0, 1]. After normalization, all data are subsequently rescaled to [-1, 1] for training various components of MRGen, including the autoencoder, diffusion UNet, and mask condition controller. For training downstream segmentation models, images are rescaled to [0, 255] and saved in '.png' format, followed by the official preprocessing configurations of nnUNet [11] and UMamba [20].

Data augmentation. During autoencoder training, we apply random data augmentations to images with a 20% probability. These augmentations include horizontal flipping, vertical flipping, and rotations of 90°, 180°, 270°. In contrast, no data augmentations are applied during the training of the diffusion UNet and mask condition controller. Our preliminary experiments show that MRGen remains robust to uneven data distribution; we therefore do not explicitly adopt data balancing in training. For segmentation models, we adhere to the default data augmentation strategies provided by nnUNet [11] and UMamba [20].

C.2. Autofilter Pipeline

When deploying our proposed data engine, MRGen, to synthesize training data for segmentation models, we adopt the off-the-shelf SAM2-Large [23] model to perform automatic interactive segmentation on generated images, with the mask conditions as spatial prompts. Empirically, we observe that SAM2 consistently segments images based on their contours, guided by the provided spatial prompts. Concretely, it produces high-quality pseudo mask annotations for images with contours closely matching mask conditions, while performing poorly for synthesized images that deviate significantly from mask conditions. This characteristic allows our pipeline to automatically filter out samples faithful to the condition masks and discard erroneous ones, thus ensuring the quality of synthesized image-mask pairs. Here, we elaborate on more implementation details of this automatic filtering pipeline, particularly focusing on the generation of MR images that encompass multiple organs.

Specifically, we begin by defining the following thresholds: confidence threshold (τ_{conf}), IoU score threshold (τ_{IoU}), average confidence threshold ($\bar{\tau}_{\text{conf}}$), and average IoU threshold ($\bar{\tau}_{\text{IoU}}$). Both the controlling mask (${\mathcal{M}'}_t$) and the generated image (${\mathcal{I}'}_t$) are fed into SAM2. For each organ mask ${\mathcal{M}'}_t^i$ in ${\mathcal{M}'}_t$, SAM2 will output a segmentation map with a confidence score (s^i_{conf}), which is then used to calculate the IoU score (s^i_{IoU}) against ${\mathcal{M}'}_t^i$. For each generated sample (${\mathcal{I}'}_t$), it is regarded to be high-quality and aligned with the mask condition if the following conditions are satisfied: $\{s^i_{\text{IoU}} \geq \tau_{\text{IoU}}, s^i_{\text{conf}} \geq \tau_{\text{conf}} \,|\, \forall i\}$, and $\{\bar{s}_{\text{IoU}} \geq \bar{\tau}_{\text{IoU}}, \bar{s}_{\text{conf}} \geq \bar{\tau}_{\text{conf}}\}$. Otherwise, the sample will be discarded.

For each conditional mask, we synthesize 20 image candidates and select the best two that satisfy the predefined thresholds. Across all experiments, the thresholds are set as follows: $\tau_{\text{IoU}} = 0.70$, $\tau_{\text{conf}} = 0.80$, $\bar{\tau}_{\text{IoU}} = 0.80$, and $\bar{\tau}_{\text{conf}} = 0.90$.

C.3. Baselines

In this section, we introduce the implementation details of representative baselines and discuss other relevant methods by category. Concretely, we first consider the most related ones, including augmentation-based and translation-based methods.

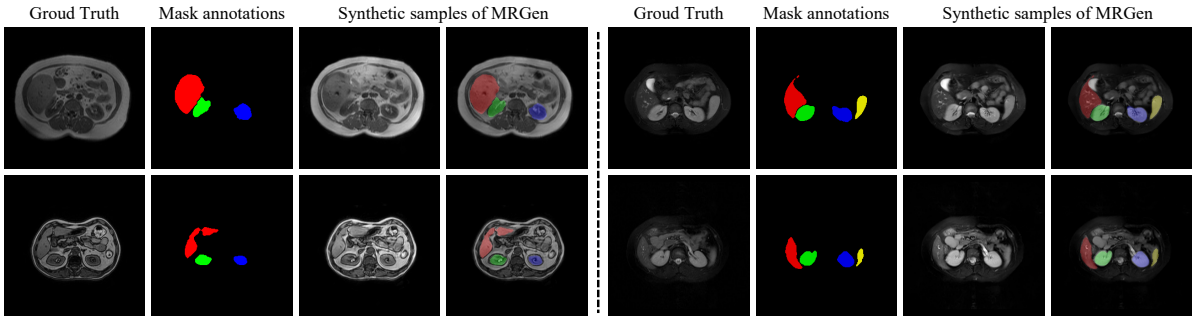
Augmentation-based methods. These approaches [5, 10, 21, 27, 29, 34] typically rely on mixing multi-domain training data or employing meticulously designed data augmentation strategies. Here, we consider the representative one, DualNorm [34]. Following its official implementation, we apply random non-linear augmentation on each source-domain image, to generate a source-dissimilar training sample, and train the dual-normalization model. All preprocessing steps, network architectures, and training strategies adhere to the official recommendations, with the exception that images are resized to 512 × 512, consistent with other methods. Notably, we evaluate DualNorm on all slices in the test set, offering a more rigorous evaluation compared to the official code, which only considers slices with segmentation annotations.

Translation-based methods. These methods [14, 15, 22, 26] are commonly inspired by CycleGAN [35]; therefore, we compare with open-source CycleGAN [35], UNSB [14], and MaskGAN [22]. We follow their official implementations and training strategies across all experimental settings. Subsequently, source-domain images are translated into the target domain and paired with the source-domain masks to create paired samples for training downstream segmentation models.

Moreover, we have also explored other approaches leveraging the progress of generative models.

Generation-based methods. Existing medical generation models [6–8, 28, 30] still struggle with complex and challenging abdominal MRI generation. For instance, MAISI [7] and Med-DDPM [6] are tailored for CT and brain MRI synthesis, respectively. To adapt to our task, we finetune MAISI [7] on our data, as a generation-based baseline.

Additionally, we consider other methods aimed at addressing our focused challenge, *i.e.*, segmenting MR images of underrepresented modalities lacking mask annotations. These include few-shot learning approaches, general-purpose segmentation models, and methods incorporating oracle inputs as performance references. Notably, these approaches, to varying degrees, rely on manually annotated target-domain segmentation masks or external datasets. Thus, they should be regarded as references only, rather than fair comparisons with the aforementioned methods and our MRGen.



(a) In-domain generation on CHAOS-MRI T1 (b) In-domain generation on CHAOS-MRI T2-SPIR

Figure 2. **Qualitative Results of In-domain Generation.**

Few-shot methods. Specifically, we compare with a few-shot nnUNet [11] (pre-trained on source-domain data and finetuned on 5% target-domain manually annotated data), as well as UniVerSeg [4] with its official implementation and checkpoint.

General segmentation models. We adopt the official code and checkpoint of TotalSegmentor-MRI [2], which has been trained on extensive manually annotated data and diverse modalities, as a strong general-purpose segmentation baseline.

Models with oracle inputs. We include SAM2-Large [23] as a reference for interactive semi-automatic segmentation, using randomly perturbed oracle boxes as prompts. To simulate the error introduced by manual intervention, the oracle boxes are randomly shifted at each corner, by up to 8% of the image resolution, following MedSAM [19]. Segmentation results are derived in a slice-by-slice and organ-by-organ manner: For each slice with mask annotations, we simulate box prompts for each annotated organ individually. Finally, we also include nnUNet [11] trained exclusively on the target-domain mask-annotated dataset ($\mathcal{D}_t$) as an oracle reference, reflecting the performance upper bound with sufficient annotated data.

D. More Experiments

In this section, we present additional experimental results to demonstrate the superiority of our proposed data engine. First, in Sec. D.1, we showcase quantitative and qualitative results of in-domain generation. Next, in Sec. D.2, we present quantitative comparisons with more baselines, further confirming the effectiveness and necessity of our proposed data engine. Then, in Sec. D.3, we present extra promising application prospects (cross-protocol generation) on paired CT and MRI datasets. Finally, in Sec. D.4, we provide extra qualitative results to validate the accuracy and flexibility of the generated outputs.

D.1. In-domain Generation

Our proposed data engine not only synthesizes images for target modalities lacking mask annotations but also maintains controllable generation capabilities within the source domains. Moreover, as presented in Table 3, downstream segmentation models trained exclusively on synthetic source-domain data can achieve performance comparable to those trained on real, manually-annotated data. This offers a feasible solution to address concerns about medical data privacy.

Dataset	Source Modality	Target Modality	$\mathcal{D}_s$		$\mathcal{D}_t$		
			$\mathcal{D}_s$	**MRGen**	$\mathcal{D}_s$	**MRGen**	$\mathcal{D}_t$
CHAOS-MRI [13]	T1	T2-SPIR	90.60	**88.14**	4.02	**67.35**	83.90
	T2-SPIR	T1	83.90	**82.06**	0.62	**57.24**	90.60

Table 3. **In-domain & Cross-domain Augmentation Results (DSC score) on Segmentation**. We compare the performance of nnUNet [11] trained on real data versus synthetic data generated by MRGen in both the source domain ($\mathcal{D}_s$) and target domain ($\mathcal{D}_t$).

Moreover, we provide visualizations of in-domain generation in Figure 2, qualitatively demonstrating that our MRGen can reliably perform controllable generation of a large number of samples within the training domain with mask annotations.

D.2. More Quantitative Results

In this section, we compare MRGen with additional baseline methods on two typical cross-modal dataset pairs from MRGen-DB by evaluating the performance of downstream segmentation models, as detailed in the main text. Concretely, for both translation-based and generation-based methods, we assess the performance of nnUNet [11] trained on data generated by these methods. As depicted in Table 4, we further analyze the relevant baselines by category, as follows.

Dataset	Source Modality	Target Modality	DualNorm	nnUNet							UniVerSeg	TS-MRI	Oracle Box	SAM2	nnUNet $\mathcal{D}_t$
				$\mathcal{D}_s$	**MRGen**	CycleGAN	UNSB	MaskGAN	MAISI	Few-shot					
CHAOS-MRI	T1	T2-SPIR	14.00	6.90	**66.18**	7.58	14.03	<u>32.73</u>	3.34	52.00	48.91	80.64	45.45	53.12	83.90
	T2-SPIR	T1	<u>12.50</u>	0.80	**58.10**	1.38	6.44	1.89	3.11	53.82	52.79	77.09	43.48	51.94	90.60
MSD-Prostate	T2	ADC	1.43	5.52	**57.83**	40.92	<u>52.99</u>	29.14	9.15	20.28	0.0	0.0	61.50	65.39	82.35
	ADC	T2	12.94	22.20	**61.95**	<u>57.06</u>	38.39	5.98	6.94	29.38	53.90	0.0	61.07	66.40	89.80
Average DSC score			10.22	8.86	**61.02**	26.74	<u>27.96</u>	17.44	5.64	38.87	38.90	39.43	52.88	59.21	86.66

Table 4. **More Quantitative Results (DSC score) on Segmentation.** The best and second-best performances are **bolded** and <u>underlined</u>, respectively. Notably, the results marked with a gray background indicate that the corresponding methods may have accessed target-modality annotated data during extensive training (*e.g.*, UniVerSeg, TotalSegmentor-MRI (TS-MRI)), utilized oracle inputs as prompts (*e.g.*, Oracle Box, SAM2), or even been directly trained on target-modality annotated data (*e.g.*, nnUNet (Few-shot), nnUNet ($\mathcal{D}_t$)). Consequently, these approaches do not represent a fully fair comparison with others, and are primarily included as performance references.

Augmentation-based methods. Limited to relying on carefully crafted augmentation strategies, DualNorm [34] fails to model nonlinear visual discrepancies among distinct modalities, leading to poor cross-modality segmentation performance.

Translation-based methods. While CycleGAN [35], UNSB [14], and MaskGAN [22] excel at contour preservation, they often suffer from model collapse when learning complex modality transformations, resulting in suboptimal performance.

Generation-based models. Despite finetuned on our dataset, the performance of MAISI [7] is still poor, which we attribute to its lack of **modality-conditioning** capability. This limitation hinders its ability to support **cross-modality generation**, and consequently, makes it struggle to synthesize target-domain samples for training segmentation models.

Few-shot methods. While few-shot nnUNet [11] and UniverSeg [4] benefit from partial target-domain annotations, MRGen-boosted models outperform without requiring any such annotations, showcasing practical feasibility in clinical scenarios.

General segmentation models. TotalSegmentor-MRI [2] works well on certain datasets/modalities (**likely already included during training**), but it still performs poorly or even fails on others. This significantly limits its practicality in complex clinical scenarios, especially when dealing with underrepresented modalities with diverse imaging characteristics.

Models with oracle inputs. Although SAM2 [23] with perturbed oracle boxes as prompts exhibits impressive zero-shot segmentation capabilities, our MRGen-boosted models still outperform it, trailing only the oracle nnUNet trained directly on target-domain annotated data. Moreover, as a semi-automatic method, SAM2's reliance on high-quality spatial prompts and manual intervention limits its scalability and applicability, while MRGen offers a fully automated, end-to-end solution.

Overall, MRGen provides a robust, fully automated approach for challenging cross-modality segmentation by producing high-quality synthetic data, with no need for any target-domain mask annotations and proving highly suitable for clinical applications. For computational efficiency, we primarily focus on comparing MRGen with some representative baselines, DualNorm [34], CycleGAN [35] and UNSB [14], across more dataset pairs in the main text for a comprehensive evaluation.

Source Domain	Target Domain	DualNorm	nnUNet			UMamba		
			$\mathcal{D}_s$	CycleGAN	**MRGen**	$\mathcal{D}_s$	CycleGAN	**MRGen**
AMOS22 (CT)	CHAOS (T2)	19.78	0.11	6.75	<u>22.50</u>	0.05	8.06	**26.73**
AMOS22 (CT)	CHAOS (T1)	16.09	8.88	52.49	<u>56.23</u>	3.19	43.21	**60.53**
MSD-Liver (CT)	CHAOS (T2)	1.58	3.12	10.14	<u>38.67</u>	1.65	11.06	**40.93**

Table 5. **Quantitative Results (DSC score) on Cross-protocol settings (from CT to MRI).**

D.3. Extension to More Modalities

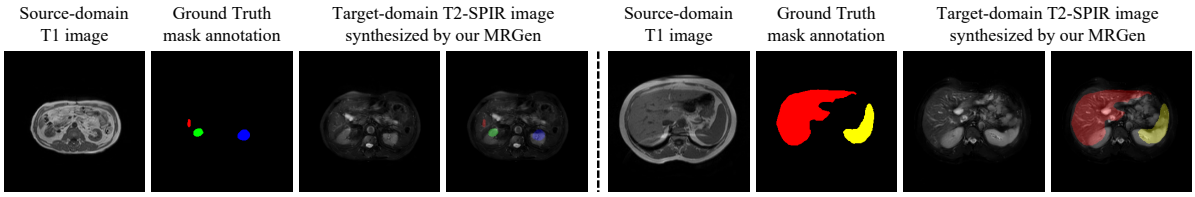
Considering that evaluation on truly rare modalities is difficult due to limited ground truth annotations, we simulate such scenarios by restricting models from accessing target-modality labels in our experiments. Here, we also explore cross-modality synthesis (from CT to MRI) with AMOS22 [12], MSD-Liver [3], and CHAOS-MRI [13] datasets to further demonstrate MRGen's potential for broader cross-protocol generation, as depicted in Table 5.

D.4. More Qualitative Results

In this section, we provide qualitative visualizations of more datasets, covering both image generation and segmentation.

Image generation. We present extra visualizations of controllable generation on target modalities lacking mask annotations in Figure 5, which demonstrate that our MRGen can effectively generate high-quality samples based on masks across various datasets and modalities, facilitating the training of downstream segmentation models towards these challenging scenarios.

Image segmentation. As presented in Figure 6, we provide more visualizations of segmentation models trained using synthetic data on modalities that originally lack mask annotations. This validates that the samples generated by MRGen can effectively assist in training segmentation models, achieving impressive performance in previously unannotated scenarios.



(a) MRGen stuggles with extremely small organ masks

(b) MRGen occasionally outputs false-negative samples

Figure 3. **Failure Cases Analysis.** Our proposed MRGen is not without limitations: (a) it may struggle to handle extremely small organ masks; (b) it occasionally produces false-negative samples, such as the unexpected synthesis of kidneys in the given example.

E. Limitations & Future Works

E.1. Limitations

Our proposed data engine, MRGen, is not without its limitations. Specifically, MRGen encounters difficulties when generating conditioned on extremely small organ masks and occasionally produces false-negative samples.

Extremely small organ masks. The morphology of the same organ, such as the *liver* or *spleen*, can vary significantly across different slices of a 3D volume, resulting in significant variability in their corresponding masks. Furthermore, the distribution of these masks is often imbalanced, with extremely small masks being relatively rare. When generating in the latent space, these masks are further downsampled, leading to unstable generation quality, as depicted in Figure 3 (a). A feasible solution to mitigate this issue is to increase the amount of data with mask annotations, thereby improving the model's robustness.

False-negative samples. Another challenge arises from the varying number of organs on each slice. For instance, one slice may contain *liver*, *kidneys*, and *spleen*, while another may include only *liver* and *spleen*. This variability causes MRGen to occasionally generate targets not specified in the mask condition. As depicted in Figure 3 (b), *kidneys* are unexpectedly synthesized, despite not being included in the mask conditions, leading to false negatives during the training of downstream segmentation networks. A feasible solution is to design a more robust data filtering pipeline to filter false-negative samples, and simple manual selection can also serve as a quick and effective method to remove the non-compliant samples.

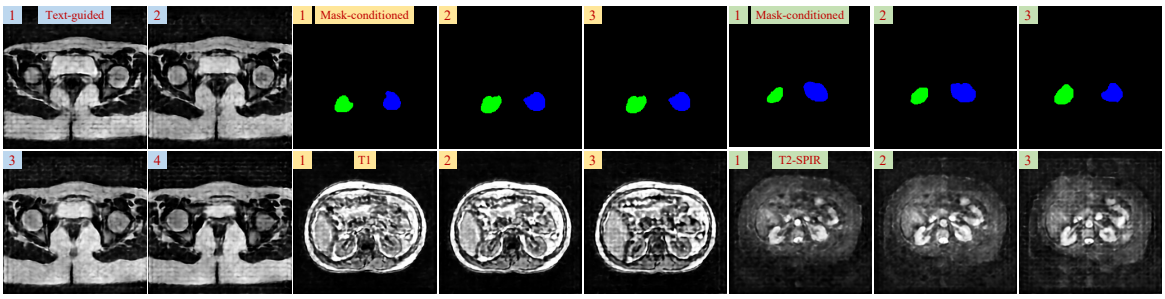


Figure 4. **The validation of 3D extension feasibility of MRGen on** $256 \times 256 \times 16$ **volumes.**

E.2. Future Works

Due to limited computational resources, we validate our data engine on 2D slices, with trained segmentation models able to process 3D volumes slice-by-slice. However, our idea can be seamlessly extended to 3D volume generation with more computing in the future to further advance cross-modality segmentation performance. Here, we provide a preliminary validation on $256 \times 256 \times 16$ volumes, as depicted in Figure 4. While the results are not fully optimized due to limited computations, they already demonstrate promising **inter-slice consistency**, indicating the feasibility of extending MRGen to 3D synthesis.

Moreover, to address the aforementioned limitations of MRGen, we propose several directions for future improvement: (i) Constructing more comprehensive and richly annotated datasets, such as incorporating more annotated MRI data, to enhance the model's ability to effectively utilize mask conditions; (ii) Designing finer-grained and efficient generative model architectures to improve generation efficiency and accuracy, particularly for small-volume organs; and (iii) Developing a more robust data filtering pipeline to reliably select high-quality samples that meet the requirements of downstream tasks.

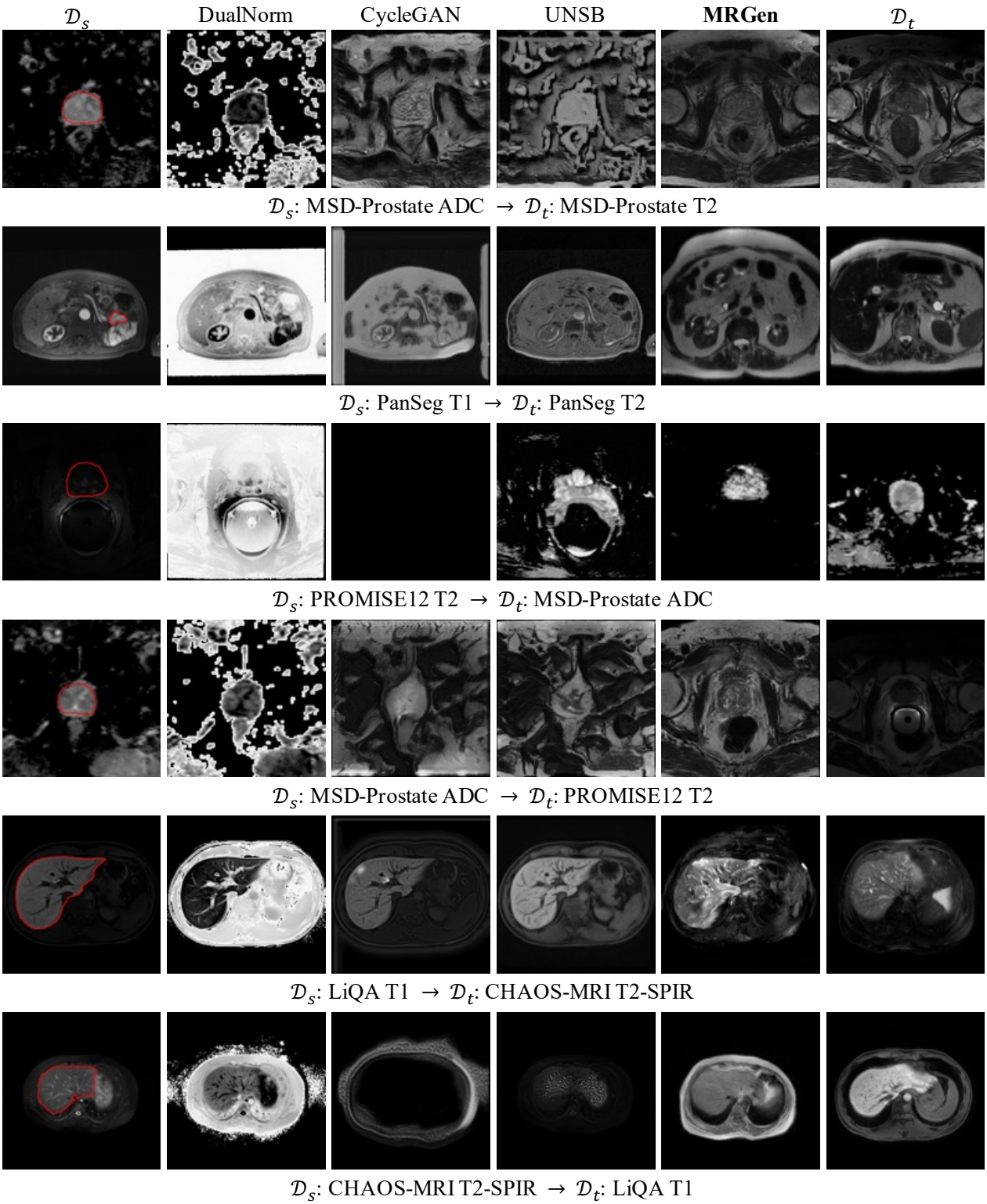


Figure 5. **More Qualitative Results of Controllable Generation.** We present images from source domains ($\mathcal{D}_s$) and target domains ($\mathcal{D}_t$) for reference. Here, specific organs are contoured with colors: prostate in MSD-Prostate and PROMISE12 datasets, and pancreas in PanSeg dataset, and liver in LiQA and CHAOS-MRI datasets.

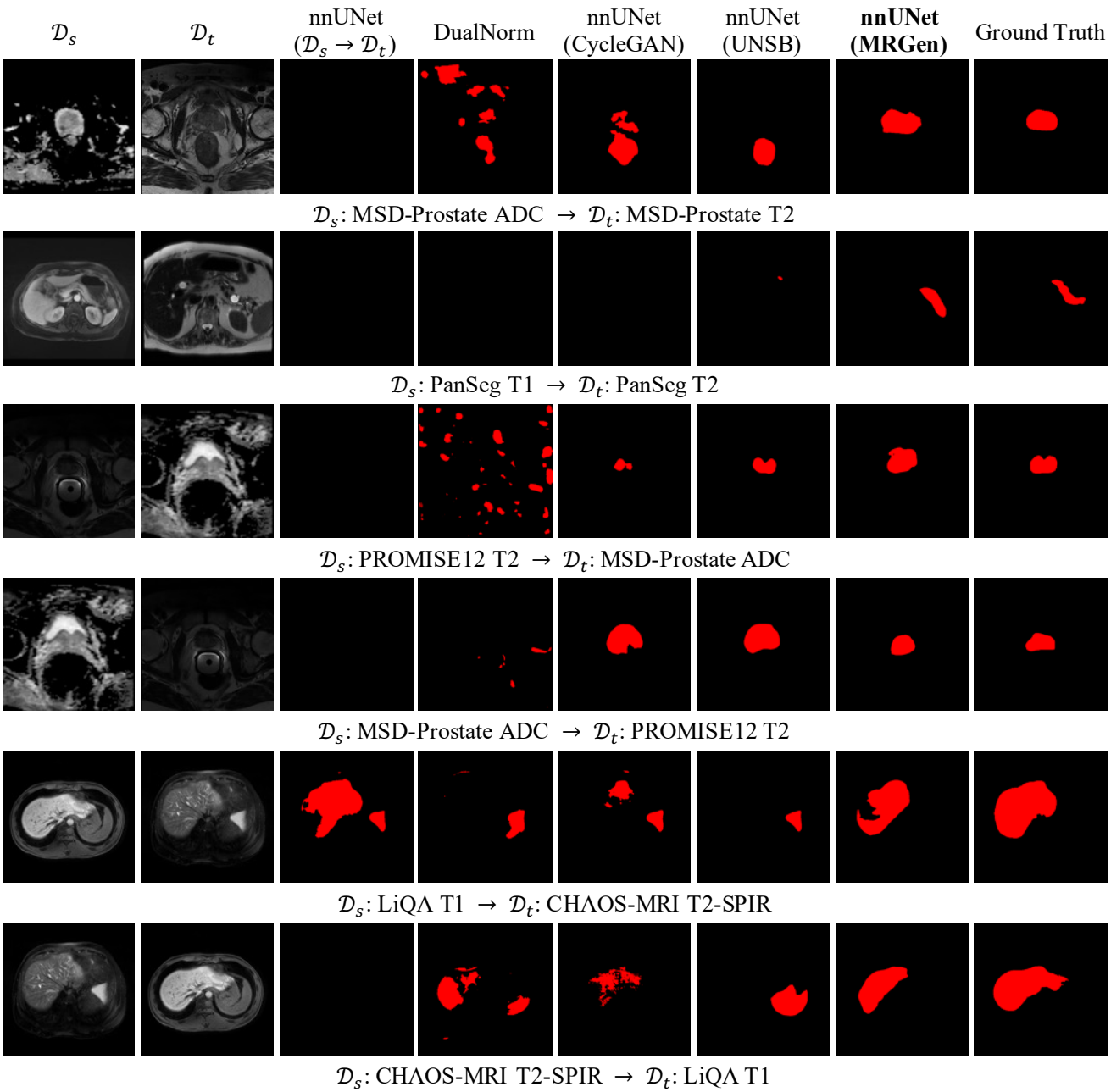


Figure 6. **More Qualitative Results on Segmentation towards Unannotated Modalities.** Significant imaging differences between source-domain ($\mathcal{D}_s$) and target-domain ($\mathcal{D}_t$) make segmentation on target domains ($\mathcal{D}_t$) extremely challenging. Here, specific organs are highlighted with colors: prostate in MSD-Prostate and PROMISE12, pancreas in PanSeg, and liver in LiQA and CHAOS-MRI datasets.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Tugba Akinci D'Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiss, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. Totalsegmentator mri: Robust sequence-independent segmentation of multiple anatomic structures in mri. *Radiology*, 314(2):e241613, 2025. 6, 7

[3] 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. 4, 7

[4] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the International Conference on Computer Vision*, pages 21438–21451, 2023. 6, 7

[5] Ziyang Chen, Yongsheng Pan, Yiwen Ye, Hengfei Cui, and Yong Xia. Treasure in distribution: a domain randomization based multi-source domain generalization for 2d medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 89–99, 2023. 5

[6] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 2024. 5

[7] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *Winter Conference on Applications of Computer Vision*, 2025. 5, 7

[8] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *Proceedings of the European Conference on Computer Vision*, pages 126–143, 2024. 5

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 2

[10] Shishuai Hu, Zehui Liao, and Yong Xia. Devil is in channels: Contrastive single domain generalization for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 14–23, 2023. 5

[11] 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. 5, 6, 7

[12] 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 *Advances in Neural Information Processing Systems*, pages 36722–36732, 2022. 7

[13] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 2021. 4, 6, 7

[14] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *Proceedings of the International Conference on Learning Representations*, 2024. 5, 7

[15] Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Winter Conference on Applications of Computer Vision*, 2024. 5

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014. 2

[17] 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. 4

[18] Yuanye Liu, Zheyao Gao, Nannan Shi, Fuping Wu, Yuxin Shi, Qingchao Chen, and Xiahai Zhuang. Merit: Multi-view evidential learning for reliable and interpretable liver fibrosis staging. *Medical Image Analysis*, 2025. 4

[19] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:1–9, 2024. 6

[20] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 5

[21] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022. 5

[22] Vu Minh Hieu Phan, Zhibin Liao, Johan W Verjans, and Minh Son To. Structure-preserving synthesis: Maskgan for unpaired mr-ct translation. In *Medical Image Computing and Computer-Assisted Intervention*, 2023. 5, 7

[23] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *Proceedings of the International Conference on Learning Representations*, 2025. 5, 6, 7

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. 2

[26] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021. 5

[27] Zixian Su, Kai Yao, Xi Yang, Kaizhu Huang, Qiufeng Wang, and Jie Sun. Rethinking data augmentation for single-source domain generalization in medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2366–2374, 2023. 5

[28] Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fei Liu, Zixing Zou, Yuanxu Gao, Lei Yang, et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, pages 1–9, 2024. 5

[29] Yanwu Xu, Shaoan Xie, Maxwell Reynolds, Matthew Ragoza, Mingming Gong, and Kayhan Batmanghelich. Adversarial consistency for single domain generalization in medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2022. 5

[30] Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11502–11512, 2024. 5

[31] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. 2

[32] Zheyuan Zhang, Elif Keles, Gorkem Durak, Yavuz Taktak, Onkar Susladkar, Vandan Gorade, Debesh Jha, Asli C Ormeci, Alpay Medetalibeyoglu, Lanhong Yao, et al. Large-scale multi-center ct and mri segmentation of pancreas with deep learning. *Medical Image Analysis*, 2025. 4

[33] 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 prompts. *arXiv preprint arXiv:2312.17183*, 2023. 3

[34] Ziqi Zhou, Lei Qi, Xin Yang, Dong Ni, and Yinghuan Shi. Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20856–20865, 2022. 5, 7

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision*, 2017. 5, 7