ViCTr: Vital Consistency Transfer for Pathology Aware Image Synthesis

Supplementary Material

A. Quantitative

Figure 7 provides a quantitative comparison of segmentation models trained with various datasets, visualized using violin plots for organs such as the aorta, left kidney, right kidney, right adrenal gland, prostate, postcava, left adrenal gland, gallbladder, and esophagus. Models trained with our synthetic data generated by **ViCTr** show improved performance over those trained with default datasets, standard data augmentation, and synthetic data generated by standard fine-tuning methods. This further validates the efficacy of our approach in enhancing segmentation tasks.

Figure 9 showcases the capability of **ViCTr** to control the severity of synthetic cirrhosis in generated images. We compare the severity levels mild, moderate, and severe between real cirrhotic images and our synthetic counterparts for both male and female subjects. The synthetic images accurately reflect the specified severity levels, and in some cases, they rank better in visual assessments than real images. This highlights the potential of our method for generating controlled pathological variations for training and diagnostic purposes.

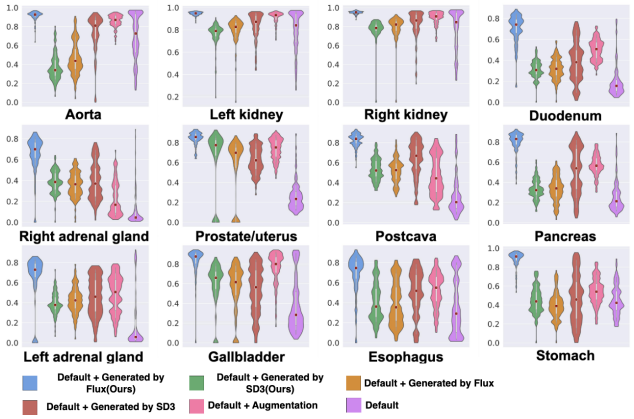

Figure 7. Segmentation Performance Comparison Using Violin Plots.

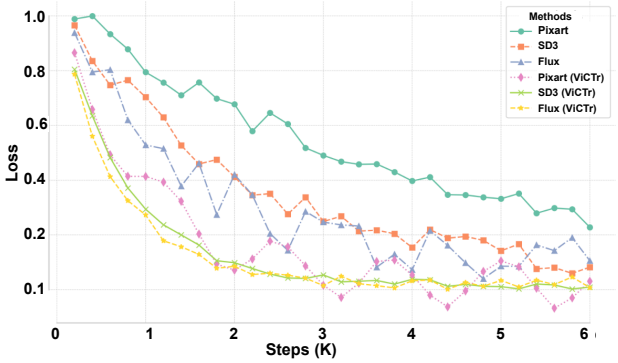

Figure 8. Convergence of models.

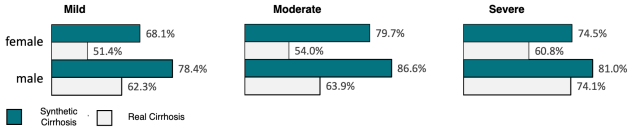

Figure 9. Comparison of severity levels (mild, moderate, severe) between real cirrhotic images and synthetic cirrhotic images generated by ViCTr for male and female subjects.

Model	Vanilla FineTuning	ViCTr (Ours)
Stable Diffusion	23.11 / 84.72	25.27 / 86.72
Stable Diffusion-XL	24.34 / 85.72	26.78 / 87.93
Stable Diffusion-3	26.44 / 87.51	28.92 / 90.37
Pixart	26.32 / 88.21	31.09 / 91.33
Flux	27.51 / 90.21	33.33 / 94.05

Table 5. PSNR / SSIM (%) comparison between Vanilla FineTuning and our ViCTr across diffusion models on CirrMRI600+.

Learning Efficiency of ViCTr-Enhanced Models: Figure 8 presents a comprehensive analysis of model convergence across 30 training steps, comparing **ViCTr** against baseline approaches. The results demonstrate **ViCTr's** superior convergence characteristics and learning efficiency across multiple state-of-the-art architectures (Pixart, SD3, and Flux—using standard vanilla fine-tuning). Lower loss values indicate better convergence, with a steeper decline in the early steps suggesting faster learning. The baseline models (Pixart, SD3, and Flux) trained with vanilla fine-tuning show a gradual decrease in loss but maintain relatively higher loss values throughout the training steps. For example, Pixart has the slowest convergence, with its loss remaining comparatively high even after 30 steps. In contrast, the **ViCTr**-enhanced models demonstrate much faster convergence rates and achieve significantly lower loss values. The consistent performance improvements across different architectures (Pixart, SD3, and Flux) further demonstrate the versatility and generalizability of our approach, establishing **ViCTr** as a powerful framework for advancing medical image synthesis.

Additional Segmentation Results:

We present extended visual results showcasing segmentation performance on complex organs such as the spleen,

liver, aorta, and stomach. As depicted in Figure 11, our method, which leverages synthetic data generated via the Flux (**ViCtr**) framework, demonstrates superior alignment with ground truth (GT) segmentation. Notably, the quality and consistency of the predicted masks across all four organ classes are on par with GT annotations. These results highlight the efficacy of our approach in capturing intricate organ structures with high precision and robustness.

Modality Translation Results on CirrMRI600+

Experimental Setup

To evaluate cross-modality translation performance, see in Table 5, we conducted experiments using the paired T1–T2 volumes from the CirrMRI600+ dataset. The goal was to synthesize target modality (T2-weighted) images conditioned on anatomical features from the source modality (T1-weighted) using text-based prompts such as *"Generate the pathology on T2-weighted MRI"*.

We assessed both structural preservation and pathological fidelity of the translated outputs. Quantitative evaluation was carried out using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), comparing synthesized T2 volumes against ground truth.

Results

Our proposed ViCTr framework consistently outperformed baseline diffusion-based models across all metrics, demonstrating superior anatomical consistency and modality-specific detail reconstruction. These findings emphasize the potential of ViCTr in downstream clinical applications such as modality harmonization, synthetic augmentation, and diagnostic support.

B. Qualitative

This Figure 10 presents a additional visual results of synthetic MRI images generated using **ViCTr**.

C. Training and Implementation Details

Pre-training. We pre-trained **ViCTr** Stage-1 using a rectified flow strategy, with the maximum diffusion steps set to 100. The Atlas-8K dataset was used as the foundational dataset, and training was performed at an image resolution of 256×256. We employed a batch size of 8, with gradient accumulation over 8 steps. Optimization was carried out using the Adam optimizer with an initial learning rate of 1×10^{-5}, managed by a cosine annealing scheduler to ensure a smooth decay of the learning rate over time. The pre-training phase was conducted on 4 nodes, each equipped with 8 Nvidia A100 GPUs 80GB each, and completed in approximately 52 hours.

Fine-tuning. For fine-tuning, we initialized **ViCTr** Stage-2 with the pre-trained weights from Stage-1 and configured it for the downstream tasks of CT, MRI, and pathological image generation. Fine-tuning was carried out at a 256×256 resolution, using a batch size of 4 with gradient accumulation over 12 steps. The Adam optimizer was used but with a higher initial learning rate of 1×10^{-4}, and a cosine learning rate scheduler for adaptive adjustment throughout training. Fine-tuning was conducted on a 2-node setup, each equipped with 8 Nvidia A100 GPUs 80GB each. Given Table below shows

Training H-Prameters	Values
Learning Rate	1.00E-04
Gradient Accumalation Steps	8
Batch Size Per GPU	2
Optimizer	AdamW
Lr-Schedular	Cosine
Epochs	40
Noise Schedular	FlowMatching
Diffusion Steps	100
Training Precision	BFloat16
GPUs	8 x 8 A100
Text Encoders	T5-XXXL
Time Embedding Size	512
Gradient Clipping	2.5
Max Text Length	200
Embedding Size	4096
CFG Scale	10.5
Positional Encodings	RoPE

Table 6. Hyper-parameters used to train models

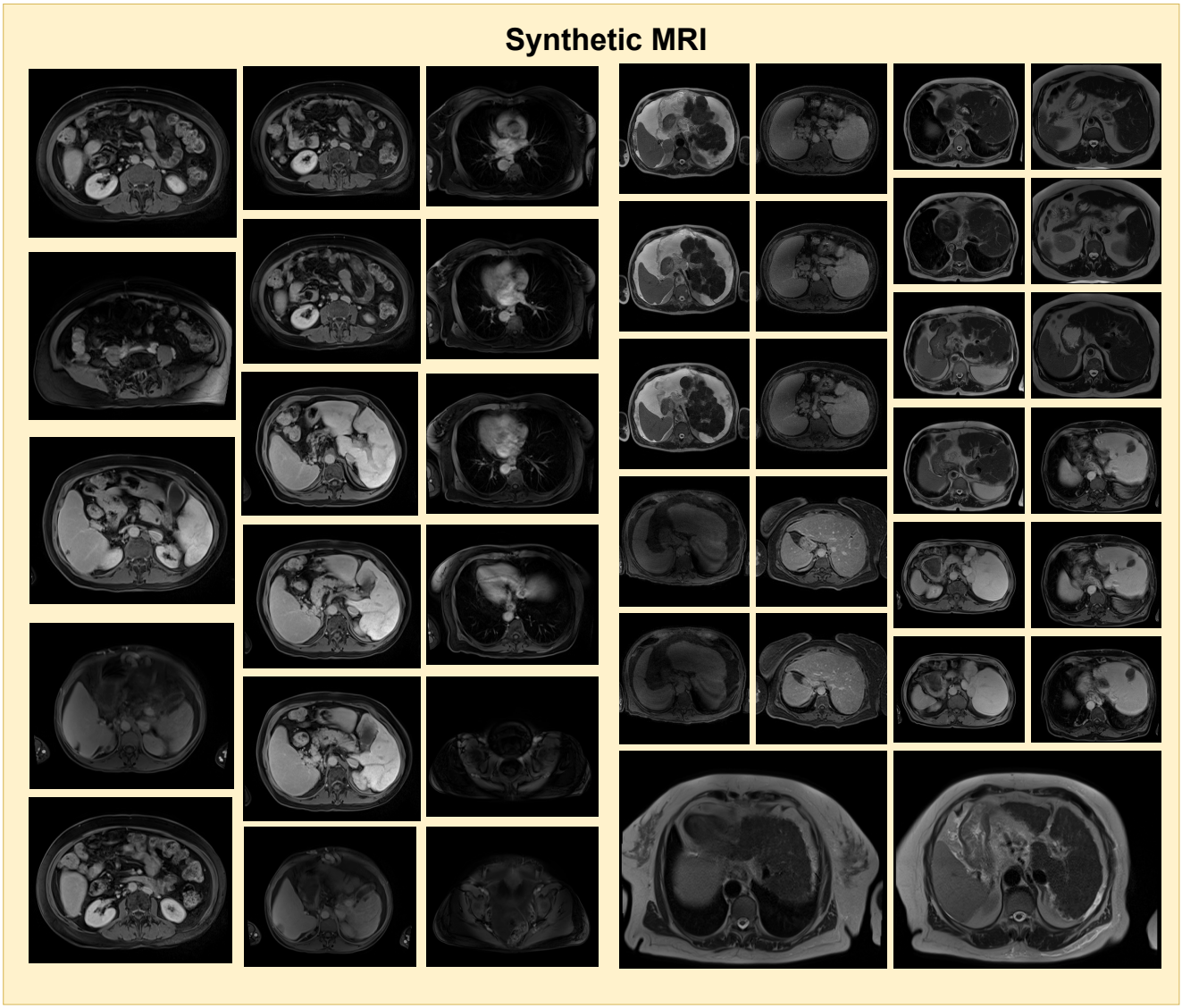

Figure 10. Synthetically generated MRI images using **ViCTr**.

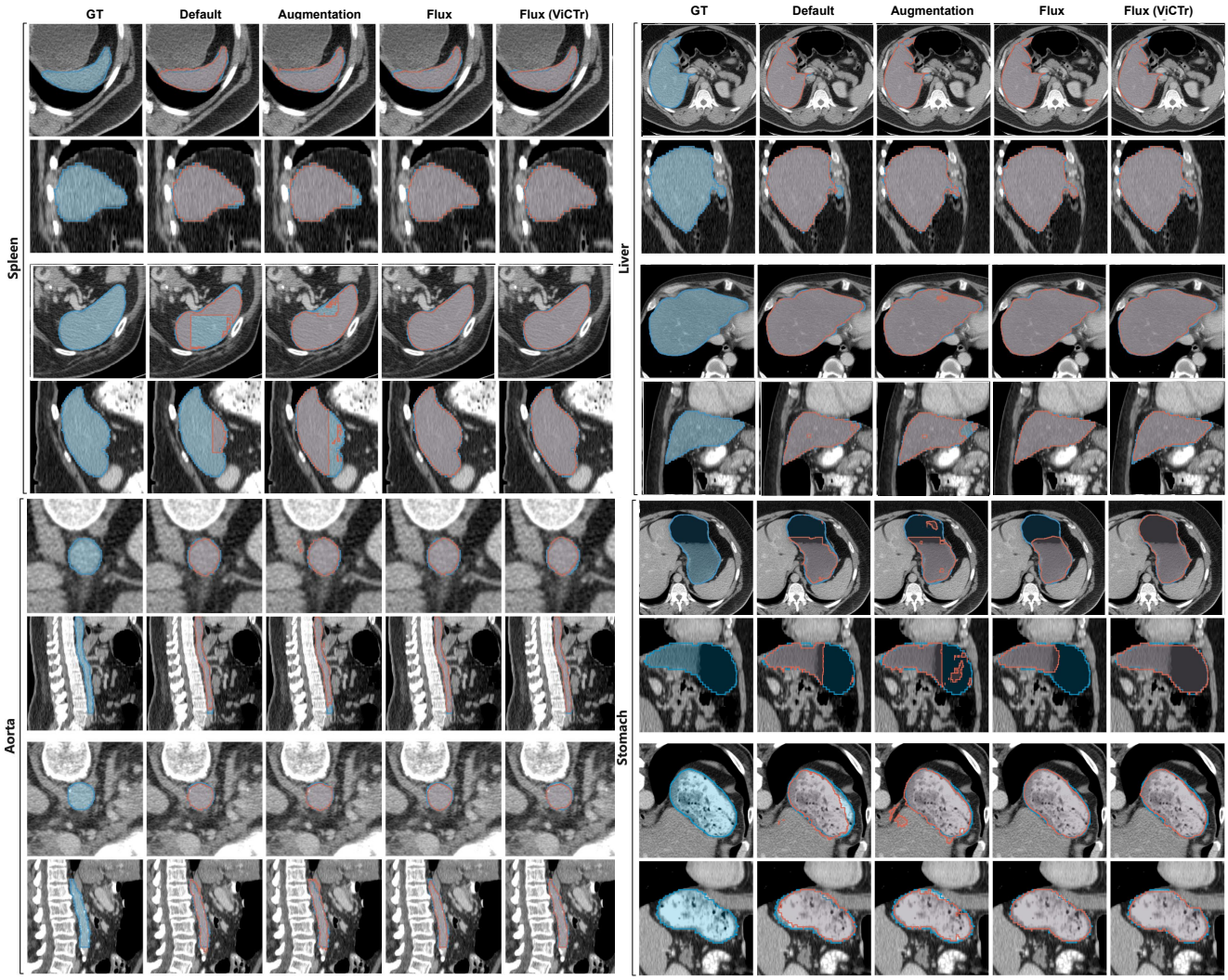

Figure 11. Segmentation results for comparison across various methods