SD-FSMIS: Adapting Stable Diffusion for Few-Shot Medical Image Segmentation

Supplementary Material

1. Implementation Details.

This section provides comprehensive details of our experimental setup to ensure full reproducibility. Our code is implemented based on RPT [10] and DiffewS [9].

Framework and Model Configuration. Our framework is built upon the publicly available Stable Diffusion v1.5 model. All input images are resized to a resolution of 256×256 pixels, maintaining consistency with the protocols of prior FSMIS methods. The vision encoder used within our Visual-to-Textual Condition Translator (VTCT) module is a pre-trained DINOv2-small (DINOv2-s) model.

Training Data and Supervision. We do not use any ground-truth segmentation masks for training. Instead, we strictly follow the self-supervised strategy proposed in ADNet [3] to generate pseudo-masks, which serve as the sole supervisory signal. To enhance model robustness and prevent overfitting, we adopt the same data augmentation strategy as RPT [10]. Both support and query images undergo random geometric transformations (including rotation, scaling, and translation) and elastic deformations.

Evaluation Protocol. We evaluate our model following the evaluation protocol implemented in RPT. Specifically, we first select the medical image volume of a single patient from the validation fold of five-fold cross-validation as the support volume (support set), which is then excluded from the validation fold. We leverage this support volume as the support information to perform segmentation on the remaining patients in the validation fold (query set). During the segmentation process, both the support volume and each query volume are split into three consecutive sub-volumes. The middle slice within each sub-volume of the support volume is utilized to segment all slices in the corresponding sub-volume of the query volume.

Training Hyperparameters. The training process is conducted using the AdamW optimizer with a weight decay value set to 1e-2. The batch size is fixed at 1, and the training is reproducible with a random seed of 42. To stabilize the training, gradient clipping is applied to all parameters with a maximum norm of 1.0. For the U-Net backbone, we employ a fine-tuning strategy with a relatively low learning rate of 1e-5. In contrast, the trainable MLP layers use a higher learning rate of 5e-5 to facilitate faster convergence. Regarding the diffusion process, we adopt a one-step DDIM scheduler, following the exact configuration specified in DiffewS [9]. Specifically, during the image generation phase, the mask is generated in a single step starting from t=999.

Hardware and Training Environment. All experiments were conducted on a single NVIDIA A6000 GPU with 48GB of VRAM. The model training for each fold consists of 15,000 iterations, which takes approximately 6 hours to complete. The training process occupies about 18GB of GPU memory.

2. More Experiments

2.1. Validation of VAE Reconstruction Capability

A fundamental premise of our work is that the Stable Diffusion's pre-trained Variational Autoencoder (VAE) can effectively compress medical images into a meaningful latent space. To validate this, we conducted a direct reconstruction experiment, as the fidelity of reconstruction directly reflects the richness of the encoded visual features. We passed medical images and their corresponding ground-truth masks through the VAE's encoder and then its decoder. The quality of the reconstructed outputs was quantitatively measured using Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and the Structural Similarity Index (SSIM). The results, presented in Tab. 1, show very low MSE and high PSNR/SSIM values for both the images and their masks. This indicates a high-fidelity reconstruction, confirming that the VAE's latent space effectively captures the essential structural and textural features of medical anatomy. This provides a robust and reliable feature foundation upon which our adaptation modules can successfully operate.

Table 1. Quantification of VAE reconstruction quality on Abd-MRI and Abd-CT.

Dataset	Type	MSE↓	PSNR↑	SSIM↑
Abd-MRI	Image	0.0005	34.1592	0.9108
	Maks	0.0007	32.2890	0.9597
Abd-CT	Image	0.0020	27.4889	0.8172
	Maks	0.0009	32.0274	0.9461

2.2. Results under Setting 2 of CD-FSMIS

We supplement the experiments on Setting 2 of Cross-Domain Few-Shot Medical Image Segmentation (CD-FSMIS), where the results of other comparative methods are directly adopted from DIFD [2]. As presented in Tab. 2,

Table 2. Quantitative comparison (in Dice score %) of different cross-domain methods under setting 2. The best value is shown in bold font, and the second best value is underlined.

Method	Ref.	Abd-CT → MRI					Abd-MRI → CT				
		Spleen	Liver	LK	RK	Mean	Spleen	Liver	LK	RK	Mean
PANet [7]	ICCV'19	33.57	31.93	27.10	32.08	31.17	28.12	41.78	16.72	20.78	26.85
SSL-ALPNet [5]	ECCV'20	51.12	47.75	44.34	50.23	48.36	34.89	54.37	30.06	33.91	38.31
RPT [10]	MICCAI'23	52.70	50.29	40.36	56.21	49.89	48.25	53.76	38.64	45.78	46.61
DR-Adapter [6]	CVPR'24	53.66	60.06	67.01	70.28	62.75	54.43	62.52	54.15	40.81	52.98
IFA$_{T=3}$ [4]	CVPR'24	56.14	63.36	71.58	73.75	66.21	55.31	68.11	51.23	<u>46.04</u>	55.17
DIFD [2]	TMI'25	<u>57.41</u>	<u>66.31</u>	<u>75.17</u>	<u>77.64</u>	<u>69.13</u>	<u>57.02</u>	<u>74.08</u>	<u>58.18</u>	42.45	<u>57.93</u>
Ours	—	**76.08**	**72.27**	**84.86**	**88.95**	**80.54**	**77.82**	**82.97**	**71.22**	**68.07**	**74.82**

Table 3. Comparison on the Abd-MRI under setting 2.

Method	Spleen	Liver	LK	RK	Mean↑	HD95↓	ASSD↓	sample/s
UniverSeg	44.27	55.08	41.50	40.03	45.22	42.84	20.05	**0.0080**
MultiverSeg	61.42	70.03	64.33	70.72	66.62	54.59	20.76	0.0728
DiffewS	73.11	77.16	77.41	83.47	77.79	17.37	8.74	0.0768
Ours	**77.25**	**78.58**	**85.03**	**88.27**	**82.28**	**13.15**	**7.38**	0.0914

the performance of our method under Setting 2 is comparable to that obtained under Setting 1 in the main text, and it outperforms the method proposed in DIFD by a significant margin across all cross-domain scenarios. Specifically, for the cross-modality task of Abd-CT → Abd-MRI, our method achieves a mean Dice score of 80.54%, which represents a substantial improvement of 11.41% over the 69.13% achieved by DIFD. For the reverse cross-modality task of Abd-MRI → Abd-CT, our method yields an even larger performance gain of 16.89% compared with DIFD. These results fully demonstrate that our method possesses more stable generalization capabilities and can better tackle the cross-domain challenges in few-shot medical image segmentation, maintaining high segmentation accuracy when transferring across different medical imaging modalities.

2.3. Comparison with Universal Models

Prior work did not compare against universal models or report HD95/ASSD. We therefore conducted additional experiments under Setting 2 with an input resolution of 256.

Shown in Tab. 3 and Tab. 4, SD-FSMIS significantly outperforms universal models. On Abd-CT, it exceeds UniverSeg [1] and MultiverSeg [8] by +46.05% and +21.43% in mean Dice, respectively. On Abd-MRI, the gains are +37.06% and +15.66%. Universal models often fail to distinguish visually similar background tissues, leading to confused masks and high HD95, whereas our method produces more accurate boundaries.

Efficiency. Our method adopts single-step denoising, directly generating the segmentation from timestep $t = 999$, which reduces inference cost. The resulting inference time is 0.09s per image, remaining within the real-time range, although slower than UniverSeg and MultiverSeg. Importantly, this minor latency increase brings substantial gains in accuracy and cross-domain robustness, which is the core contribution of our work. In medical imaging, robustness and precision are significantly more critical than minimal latency. We therefore consider this trade-off both practical and clinically reasonable.

Visualization. As illustrated in Fig. 1, we conduct 1-shot segmentation experiments on Abd-MRI and Abd-CT, and compare our method with two representative universal segmentation models: UniverSeg and MultiverSeg. For UniverSeg, under the 1-shot support set setting, the model almost fails to perform effective segmentation of the target organs. Its generated masks are randomly scattered around the target regions with no clear structural consistency, which reflects the poor adaptability of vanilla universal models to medical image segmentation tasks with limited annotated samples. MultiverSeg shows an improved performance compared to UniverSeg and can roughly localize the target organs in both Abd-MRI and Abd-CT images. However, this model still suffers from obvious limitations in fine-grained boundary segmentation: it cannot accurately distinguish the foreground target organs from the visually similar background tissues (e.g., adjacent visceral tissues and parenchyma), thus resulting in frequent under-segmentation (missing partial valid regions of target organs) and over-segmentation (erroneously including background tissues into the segmentation masks) issues.

Table 4. Comparison on the Abd-CT under setting 2.

Method	Spleen	Liver	LK	RK	Mean↑	HD95↓	ASSD↓	sample/s
UniverSeg	34.58	51.22	31.07	31.97	37.20	53.86	24.29	**0.0077**
MultiverSeg	62.39	76.76	54.19	53.93	61.82	57.63	24.36	0.0551
DiffewS	76.84	79.57	69.70	73.62	74.93	18.86	9.49	0.0732
Ours	**83.08**	**82.59**	**82.22**	**85.10**	**83.25**	**13.20**	**7.62**	0.0856

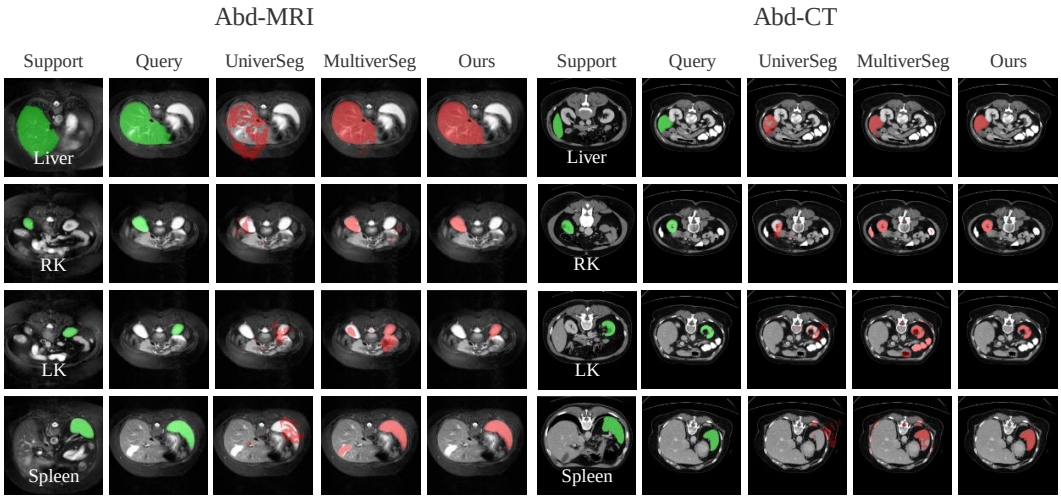

Figure 1. Qualitative comparison between our method and the universal models method on the Abd-MRI dataset and Abd-CT dataset.

In contrast, our method achieves more robust and accurate segmentation in the 1-shot scenario. It not only precisely localizes the target organs but also effectively discriminates the subtle boundary differences between the foreground organs and the background tissues with similar visual features. This superior performance is attributed to our method leveraging the pre-trained Stable Diffusion model with strong visual priors, which we adapt to the few-shot medical image segmentation task via dedicated design. This adaptation unlocks the model's powerful generalization capability and equips it with a stronger ability to capture organ-specific anatomical features and discriminate visually similar tissues.

3. Analysis and Visualization

3.1. Analysis Failure Cases

Despite the overall effectiveness of SD-FSMIS, our evaluation on the Abd-MRI dataset reveals some performance discrepancies, particularly for certain organ classes. As illustrated in Fig. 2, visual inspection of the segmentation results indicates that the model occasionally produces incomplete or over-segmented masks for the Liver. This issue appears to stem from the inherently low contrast between the liver tissue and the surrounding background, resulting in ambiguous boundaries that challenge the model's ability to distinguish between foreground and background regions.

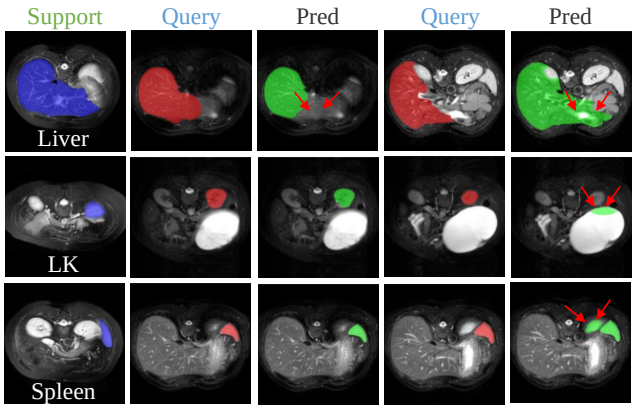

Figure 2. Visualization of failure cases on the Abd-MRI dataset.

In addition, when segmenting the Left Kidney (LK), we observe that the model's attention may be disproportionately drawn to regions with extreme saliency in a given image slice. This can lead to inconsistent performance across consecutive slices—while one slice may be segmented accurately, the subsequent slice might exhibit segmentation errors, misidentifying the target region.

Furthermore, in cases where both the Spleen and the LK

appear within the same slice and are positioned in close proximity, the model tends to merge these adjacent organs into a single segmentation output. These mis-segmentation events suggest that the spatial relationships and relative proximities between organs play a critical role in challenging the model's discriminative capacity.

These findings highlight specific challenges in medical image segmentation, where subtle contrast differences and complex anatomical interactions can lead to segmentation inaccuracies. Addressing these issues by enhancing attention mechanisms or improving boundary detection strategies may further improve the robustness of SD-FSMIS in future work.

3.2. Visualization of Training

Additionally, Fig. 3 illustrates the performance of our method during the training process on the Abd-CT dataset. Notably, even in the early stages of training (after 500 iterations), the model is able to segment simpler classes effectively, and for more complex organs such as the liver, good segmentation results are achieved as early as 5,000 iterations. These findings further underscore the powerful capabilities of diffusion models in tackling few-shot segmentation challenges.

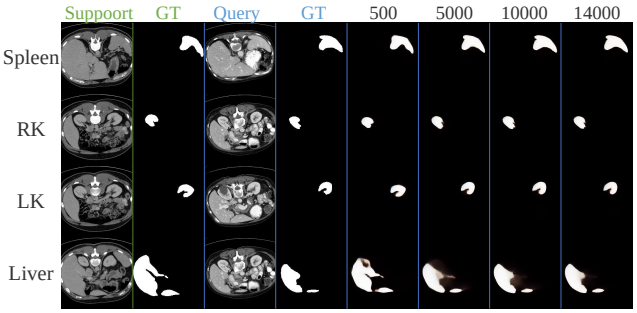

Figure 3. Visualization of the training process.

References

[1] 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 IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023. 2

[2] Ziming Cheng, Shidong Wang, Yang Long, Tao Zhou, Haofeng Zhang, and Ling Shao. Dual interspersion and flexible deployment for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025. 1, 2

[3] Stine Hansen, Srishti Gautam, Robert Jenssen, and Michael Kampffmeyer. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78:102385, 2022. 1

[4] Jiahao Nie, Yun Xing, Gongjie Zhang, Pei Yan, Aoran Xiao, Yap-Peng Tan, Alex C Kot, and Shijian Lu. Cross-domain few-shot segmentation via iterative support-query correspondence mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3380–3390, 2024. 2

[5] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *European conference on computer vision*, pages 762–780. Springer, 2020. 2

[6] Jiapeng Su, Qi Fan, Wenjie Pei, Guangming Lu, and Fanglin Chen. Domain-rectifying adapter for cross-domain few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24036–24045, 2024. 2

[7] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 2

[8] Hallee E Wong, Jose Javier Gonzalez Ortiz, John Guttag, and Adrian V Dalca. Multiverseg: scalable interactive segmentation of biomedical imaging datasets with in-context guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20966–20980, 2025. 2

[9] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 37:42672–42695, 2024. 1

[10] Yazhou Zhu, Shidong Wang, Tong Xin, and Haofeng Zhang. Few-shot medical image segmentation via a region-enhanced prototypical transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 271–280. Springer, 2023. 1, 2