SAT-Morph: Unsupervised Deformable Medical Image Registration using Vision Foundation Models with Anatomically Aware Text Prompt

Hao Xu[1], Tengfei Xue[1,2], Dongnan Liu[1], Fan Zhang[2,3], Carl-Fredrik Westin[2], Ron Kikinis[2], Lauren J. O'Donnell[2], Weidong Cai[1(✉)]

[1] University of Sydney, Sydney, Australia
[2] Harvard Medical School, Boston, USA
[3] University of Electronic Science and Technology of China, Chengdu, China
tom.cai@sydney.edu.au

Abstract. Current unsupervised deformable medical image registration methods rely on image similarity measures. However, these methods are inherently limited by the difficulty of integrating important anatomy knowledge into registration. The development of vision foundation models (e.g., Segment Anything Model (SAM)) has attracted attention for their excellent image segmentation capabilities. Medical-based SAM aligns medical text knowledge with visual knowledge, enabling precise segmentation of organs. In this study, we propose a novel approach that leverages the vision foundation model to enhance medical image registration by integrating anatomical understanding of the vision foundation model into the medical image registration model. Specifically, we propose a novel unsupervised deformable medical image registration framework, called SAT-Morph, which includes Segment Anything with Text prompt (SAT) module and mask registration module. In the SAT module, the medical vision foundation model is utilized to segment anatomical regions within both moving and fixed images according to our designed text prompts. In the mask registration module, we take these segmentation results instead of traditionally used image pairs as the input of the registration model. Compared with utilizing image pairs as input, using segmentation mask pairs incorporates anatomical knowledge and improves the registration performance. Experiments demonstrate that SAT-Morph significantly outperforms existing state-of-the-art methods on both the Abdomen CT and ACDC cardiac MRI datasets. These results illustrate the effectiveness of integrating vision foundation models into medical image registration, showing the potential way for more accurate and anatomically-aware registration. Our code is available at https://github.com/HaoXu0507/SAT-Morph/.

Keywords: Medical Image Registration · Vision Foundation Model · Text-prompted Segmentation · Multi-modal Learning

1 Introduction

Medical image registration refers to establishing the spatial correspondence between fixed images and moving images to maximize their similarity. Recently, many unsupervised deformable registration methods [2–5, 27, 12, 15, 22, 18, 6, 9] have emerged. TransMorph [5] combines the local capabilities of convolutional neural networks (CNN) [17] and the global capabilities of Transformer [23] to improve registration performance. TransMatch [6] further improves the effect by directly using the transformer's attention mechanism for image registration. DiffusionMorph [15] and its variant FSDiffReg [22] use the progressive denoising strategy of the diffusion model itself to perform progressive denoising and simultaneous registration CS-Reg [4] performs the cyclical self-training strategy to gradually refine pseudo labels.

However, these unsupervised deformable registration methods based on similarity measures can only equally weight the entire image but fail to allocate more weights to important anatomical regions. Therefore, these methods cannot integrate medical anatomy knowledge while performing registration. Currently, there is a great need for an unsupervised deformable image registration method to integrate the knowledge of medical anatomy.

The medical foundation model aligns medical text knowledge with medical image knowledge, showing promising results in various medical image tasks [28, 19, 25, 10, 14, 26, 11, 7]. In particular, Segment Anything Model (SAM) [16] has recently attracted attention in the community because of its excellent image segmentation capabilities that only require simple prompts (box, point, or mask). Various variants of SAM [21, 8, 24, 29] are constantly emerging to explore the boundaries of its capabilities, including medical image-based SAM [13]. MedSAM [20] fine-tunes SAM and integrates medical knowledge in specific fields into the segmentation model, proving SAM's effectiveness in medical image registration. As the original SAM model is based on a 2D image architecture, it is not suitable for 3D medical image segmentation. SAM-MED3D [24] builds a medical image SAM model based on 3D images. Compared with the original SAM and SAM-MED2D [8], it achieves SOTA performance by using only 10 box prompts. However, the above variants of SAM based on medical images require the assistance of vision prompts (box, point, mask, etc.), which is still time-consuming and labor-intensive. Segment anything with text prompt (SAT) model aligns the textual knowledge and visual knowledge of the structure of medical images and achieves SOTA segmentation results by using only text prompts. Although medical-based SAM can deeply understand the various anatomical structures of medical images, there is no SAM-based method dedicated for medical image registration.

In this study, we propose a SAM-driven image registration framework called SAT-Morph, including a SAT module and a mask registration module. In the SAT module, we use our designed text prompts to guide the powerful medical SAM model to segment paired registered images and generate pseudo mask labels of anatomical regions. In the mask registration module, instead of using moving and fixed image pairs as the input of the registration model, pseudo

mask labels are utilized as the input to incorporate anatomical information and improve registration accuracy. To the best of our knowledge, we are the first to use pseudo mask labels as the input of registration model. We demonstrate the superiority of our proposed framework compared to the SOTA methods on the two datasets: ACDC Cardiac MRI and Abdomen CT datasets.

Our contributions are as follows. First, we propose a novel unsupervised deformable medical image registration method driven by the vision foundation model with our designed text prompts. Second, instead of using images as the input of the registration model, we utilize pseudo mask labels as the input, which integrates anatomical knowledge and improves registration performance. Third, our framework outperforms previous methods by a significant margin on ACDC cardiac MRI and Abdomen CT datasets. The framework demonstrates a potential way for more accurate and anatomically aware registration techniques.

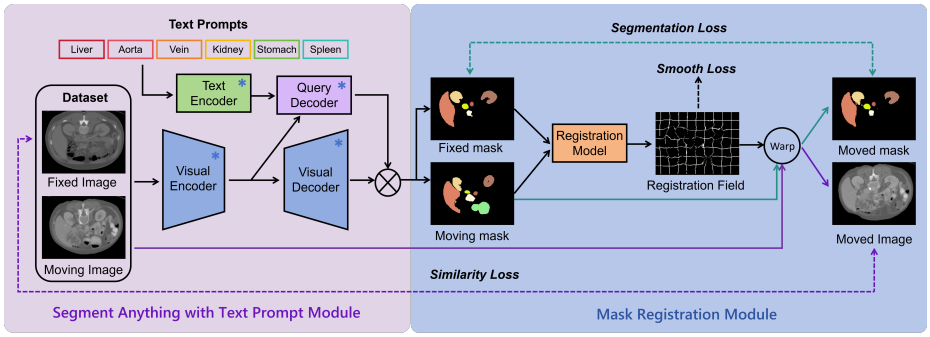

Fig. 1. The framework of SAT-Morph. Segment Anything with Text Prompt Module: Generating pseudo mask labels of image pairs according to our designed text prompts. Mask Registration Module: Taking pseudo mask labels as the input and output registration results. $*$ denotes freezing model parameters.

2 Methodology

Our framework aims to obtain a spatial transform field U for register from M to F. In the SAT module, vision foundation model SAT generates pseudo masks by segmenting fixed and moving images into anatomical regions according to our designed text prompts. In the mask registration module, to reduce the training difficulty of the registration model and improve the registration performance, we utilize the pseudo mask labels of fixed and moving images as the input of the registration model.

2.1 Segment Anything with Text Prompt (SAT) Module

We utilize SAT module to generate pseudo mask labels of fixed and moving images. The SAT module is based on a pre-trained medical segmentation foun-

dation model [29] with medical anatomical structure text as prompt, which consists of the visual encoder, visual decoder, text encoder, and query decoder. The text encoder and the query decoder take in text prompts and visual encoder and visual decoder take in medical image scans. According to the importance of organs, we generate text prompts for organs that are important for medical registration, and ignore unimportant organs. For example, for abdominal CT registration, liver and kidney are important, while colon and duodenum are unimportant. We fuse them to obtain segmentation masks according to our designed text prompts. Let a pair of fixed and moving images be $\{F, M\}$ and text prompts be $T = \{t1, ..., tn\}$:

$$F_{seg} = \theta_{SAT}(F, T), \tag{1}$$

where θ_{SAT} is the SAT segmentation model, and F_{seg} is the segmentation result of F. In the same way:

$$M_{seg} = \theta_{SAT}(M, T), \tag{2}$$

where M_{seg} is the segmentation result of M. Note that we design a set of text prompts for each dataset. Each set of text prompts can be used for all images of the entire dataset.

2.2 Mask Registration Module

We utilize the pair of $\{F_{seg}, M_{seg}\}$ as the pseudo labels of fixed and moving images. The registration model takes the pair of pseudo masks as input to compute the displacement field. Note that the field can be used as the spatial transform field not only from M_{seg} to F_{seg} but also from M to F. According to this property, we register the pseudo mask pair and the registration result can be directly applied to the image pair. Specifically, let the predicted displacement field be:

$$u = \theta_{Reg}(F_{seg}, M_{seg}), \tag{3}$$

where θ_{Reg} is the registration model. The spatial transform from the moving image to the moved image, and from the moving pseudo mask to the moved pseudo mask are as follows:

$$M' = \phi(M, u) \tag{4}$$

and

$$M'_{seg} = \phi(M_{seg}, u), \tag{5}$$

where M' is the moved image, M'_{seg} is the moved pseudo mask, and ϕ is the spatial transform function.

2.3 Loss Functions

Our loss function combines segmentation loss, similarity loss, and smooth loss.

Segmentation Loss. We calculate the segmentation loss between moved pseudo mask and fixed pseudo label to constrain the accuracy of the registration model. We take the combination of dice loss and focal loss as the segmentation loss, which is as follows:

$$L_{dice} = 1 - \frac{2\left|M'_{seg} \cap F_{seg}\right|}{\left|M'_{seg}\right| + |F_{seg}|}, \tag{6}$$

$$L_{focal} = \begin{cases} -F_{seg}(1 - M'_{seg})^{\gamma} \log M'_{seg}, & if \quad F_{seg} = 1, \\ (1 - F_{seg})(M'_{seg})^{\gamma} \log\left(1 - M'_{seg}\right), & otherwise. \end{cases} \tag{7}$$

$$L_{seg} = L_{dice} + \alpha L_{focal}, \tag{8}$$

where α is the hyper-parameter.

Similarity Loss. The mean squared error (MSE) loss between moved and fixed images is adopted as the similarity loss:

$$L_{sim} = \frac{1}{\Omega} \sum \left|M' - F\right|^2, \tag{9}$$

where Ω represents the image domain.

Smooth Loss. We utilize diffusion regularization on the spatial gradients of the deformable field as the smooth loss:

$$L_{smooth} = \sum_{p \in \Omega} || \bigtriangledown u(p)||^2, \tag{10}$$

where p denotes the voxel location. Finally, the overall loss function is as follows:

$$L = L_{seg} + \beta L_{sim} + \delta L_{smooth}, \tag{11}$$

where β and δ are the hyper-parameters.

3 Experiments and Results

3.1 Dataset and Text Prompts

Our methods are evaluated on two datasets of CT and MRI modalities: the Abdomen CT dataset and ACDC cardiac MRI dataset.

Abdomen CT Dataset. The Abdomen CT dataset contains 50 abdominal images. We randomly chose 40 images (780 pairs) for training and 10 images (45 pairs) for testing. The resolution is $192 \times 160 \times 256$ and each voxel size is $2 \times 2 \times 2$ mm. Our designed text prompts include spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland, and left adrenal gland.

ACDC Cardiac MRI Dataset. The ACDC cardiac MRI dataset contains 100 image pairs. We randomly chose 90 image pairs for training and 10 image pairs for testing, following the data split in previous works [15, 22]. The resolution is $128 \times 128 \times 32$ and each voxel size is $1.5 \times 1.5 \times 3.15$ mm. Our designed text prompts include myocardium, left ventricle, and right ventricle.

3.2 Evaluation Metrics

We use dice score coefficient (DSC) and standard deviation of the Jacobian determinant (SDlogJ) as the evaluation metrics of the experiment, which two are widely used to evaluate on image registration [15, 22, 5, 6, 2]. A higher DSC indicates that the displacement field more accurately aligns the anatomy of relevant organs between the moving and fixed images. A lower SDlogJ indicates a smoother and more consistent displacement field between the moving and fixed images.

3.3 Implementation Details

We employ the trained SAT-nano model [29] as the segmentation model in our SAT module and freeze the model parameters throughout the training and inference stages. For the registration model, we choose TransMatch [6] as our basic registration model. Following the setup of TransMatch, our method is trained using Adam with a learning rate of 0.0004 and batch size 1 for 500 epochs. Regarding hyperparameters, α is 20, β is 1 , and δ is 0.04. The experiment is performed with Pytorch (v1.10) on an NVIDIA GeForce RTX 3090 GPU machine.

3.4 Comparison Experiments

Quantitative and Qualitative Comparisons on Abdomen CT Dataset. We compare our method with seven SOTA unsupervised deformable registration methods, including SyN [1], LDDMM [3], Deeds [12], VoxelMorph [2], TransMorph [5], TransMatch [6], and CS-Reg [4]. SyN, LDDMM, and Deeds are the traditional training-free registration methods and the others are deep learning based methods. As shown in Table 1, our method exceeds SOTA methods by a large margin. Specifically, our method achieves a margin of 14.73% and 11.77% higher DSC over Deeds and CS-Reg, respectively. To better demonstrate our results, we also show organ-specific results in the supplementary materials (Fig. S1). It shows that our method achieves the best DSC in 12 out of 13 organs.

Fig. 2 depicts qualitative comparisons of our approach against Deeds, CS-Reg, and SAT-Morph w/ image. As seen, our approach achieves a more accurate registration result than competitors.

Quantitative and Qualitative Comparisons on ACDC Cardiac MRI Dataset. We compare our method with four unsupervised deformable registration methods, including VoxelMorph, VoxelMorph-Diff [9], DiffuseMorph [15], and FSDiffReg [22]. As shown in Table 2, our method exceeds VoxelMorph, VoxelMorph-DIff, DiffuseMorph, FSDiffReg by 9.8%, 11.0%, 9.2%, and 6.6% DSC, respectively. In terms of specific anatomy regions, we lead the SOTA method by 11.0% and 9.4% DSC in the LV and Myo regions. Moreover, compared with other methods, our approach achieves the smallest SDlogJ. It shows that our method generates a smoother and more consistent displacement field.

Table 1. Quantitative comparisons on Abdomen CT dataset. ↑: higher is better, and ↓: lower is better.

	Method	DSC (%) ↑	SDlogJ ↓
Comparison	SyN [1]	23.25	N/A
	LDDMM [3]	25.51	N/A
	Deeds [12]	48.99	N/A
	VoxelMorph [2]	37.67	**0.143**
	TransMorph [5]	39.03	0.254
	TransMatch [6]	42.15	0.386
	CS-Reg [4]	51.95	0.149
Ablation Study	SAT-Morph w/ image	59.39	0.974
	SAT-Morph w/ mask (Ours)	**63.72**	0.910

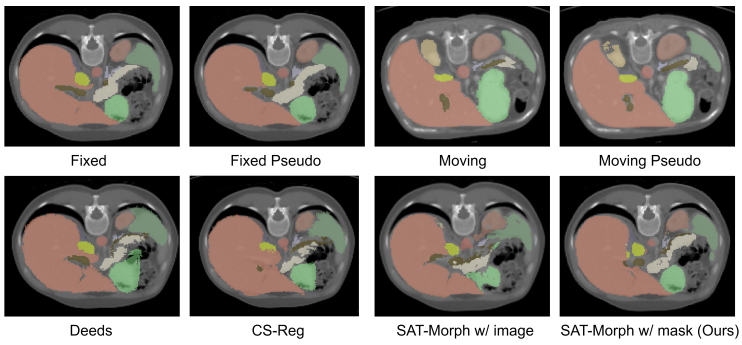

Fig. 2. Visualization of registration results for our proposed method and compared methods on Abdomen CT dataset.

The result of qualitative comparisons is shown in Fig. 3. As seen, our method has more complete and accurate registration results compared with other SOTA methods (e.g., FSDiffReg).

Table 2. Quantitative comparisons on ACDC Cardiac MRI dataset. ↑: higher is better, and ↓: lower is better. LV: left ventricle. Myo: myocardium. RV: right ventricle.

Method	DSC (%) ↑				SDlogJ ↓
	LV	Myo	RV	Overall	
VoxelMorph [2]	77.0	67.9	81.6	75.5	0.183
VoxelMorph-Diff [9]	75.5	65.9	81.5	74.3	0.182
DiffuseMorph [15]	78.3	67.8	82.1	76.1	0.178
FSDiffReg [22]	80.9	72.4	**82.7**	78.7	0.176
SAT-Morph (Ours)	**91.9**	**82.0**	81.9	**85.3**	**0.058**

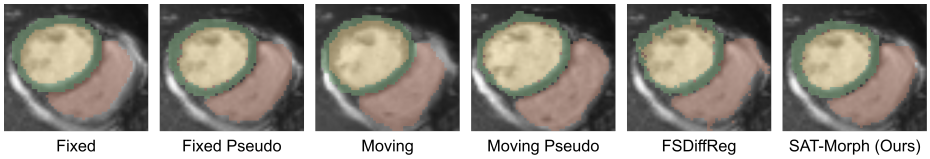

Fig. 3. Visualization of registration results for our proposed method and compared methods on ACDC Cardiac MRI dataset.

Ablation Study. We compare the results of using the fixed and moving image pair or their pseudo masks as input of the registration model. As shown in Table 2, using pseudo masks as input exceeds using the image pair by 4.33% DSC. It demonstrates that using pseudo mask pairs can better train the model to achieve better registration accuracy than using image pairs as input of the registration model.

4 Conclusion

This work proposes SAT-Morph, a novel framework that leverages vision foundation model-driven approach into unsupervised deformable medical image registration. Our framework includes SAT module and mask segmentation module. The SAT module utilizes our uniquely designed text prompts to guide the vision foundation model in generating accurate pseudo mask labels. Then, these pseudo masks are taken as inputs by the registration model, replacing the traditional image pair inputs and thereby potentially pioneering a new direction in registration methodology. We demonstrate that using pseudo masks can achieve better registration accuracy than using image pairs as inputs to the registration model. We also show significant improvements in the Abdomen CT dataset and ACDC cardiac MRI dataset, highlighting its potential to set a new way for more accurate and anatomically-aware registration.

References

1. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis **12**(1), 26–41 (2008)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging **38**(8), 1788–1800 (2019)
3. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision **61**, 139–157 (2005)
4. Bigalke, A., Hansen, L., Mok, T.C., Heinrich, M.P.: Unsupervised 3d registration through optimization-guided cyclical self-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 677–687. Springer (2023)
5. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration. Medical image analysis **82**, 102615 (2022)
6. Chen, Z., Zheng, Y., Gee, J.C.: Transmatch: A transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. IEEE Transactions on Medical Imaging **43**(1), 15–27 (2024)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
8. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
9. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical image analysis **57**, 226–236 (2019)
10. Gu, T., Liu, D., Li, Z., Cai, W.: Complex organ mask guided radiology report generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7995–8004 (2024)
11. Gu, T., Yang, K., Liu, D., Cai, W.: Lapa: Latent prompt assist model for medical visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4971–4980 (June 2024)
12. Heinrich, M.P., Maier, O., Handels, H.: Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. VISCERAL Challenge@ ISBI **1390**, 27 (2015)
13. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
14. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2607–2615 (2024)
15. Kim, B., Han, I., Ye, J.C.: Diffusemorph: unsupervised deformable image registration using diffusion model. In: European Conference on Computer Vision. pp. 347–364. Springer (2022)

16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
17. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361**(10), 1995 (1995)
18. Li, Z., Tian, L., Mok, T.C., Bai, X., Wang, P., Ge, J., Zhou, J., Lu, L., Ye, X., Yan, K., et al.: Samconvex: Fast discrete optimization for ct registration using self-supervised anatomical embedding and correlation pyramid. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 559–569. Springer (2023)
19. Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Li, Q., et al.: Differentiating chatgpt-generated and human-written medical texts: quantitative study. JMIR Medical Education **9**(1), e48904 (2023)
20. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
21. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023)
22. Qin, Y., Li, X.: Fsdiffreg: Feature-wise and score-wise diffusion-guided unsupervised deformable image registration for cardiac images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–665. Springer (2023)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
24. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d. arXiv preprint arXiv:2310.15161 (2023)
25. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36** (2024)
26. Xu, J., Lu, L., Peng, X., Pang, J., Ding, J., Yang, L., Song, H., Li, K., Sun, X., Zhang, S., et al.: Data set and benchmark (medgpteval) to evaluate responses from large language models in medicine: Evaluation development and validation. JMIR Medical Informatics **12**(1), e57674 (2024)
27. Zhang, F., Wells, W.M., O'Donnell, L.J.: Deep diffusion mri registration (ddmreg): a deep learning method for diffusion mri registration. IEEE Transactions on Medical Imaging **41**(6), 1454–1467 (2021)
28. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis p. 102996 (2023)
29. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: One model to rule them all: Towards universal segmentation for medical images with text prompts. arXiv preprint arXiv:2312.17183 (2023)