A Multimodal 3D Foundation Model for Light Sheet Fluorescence Microscopy Enables Few-Shot Segmentation, Classification, and Deblurring

Adina Scheinfeld[1,2], Haotan Zhang[3,4], Shang Mu[3], Rudolf L. M. van Herten[2,5], Lucas Stoffl[2,5], Ali Ertürk[6,7], Zhuhao Wu[3], and Johannes C. Paetzold[2,5]

[1] Tri-Institutional Program in Computational Biology & Medicine, Weill Cornell Medicine, New York, NY, USA
ads4015@med.cornell.edu
[2] Department of Radiology, Weill Cornell Medicine, New York, NY, USA
[3] Helen and Robert Appel Alzheimers Disease Research Institute, Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA
[4] Graduate Program in Physiology, Biophysics and Systems Biology, Weill Cornell Medicine, New York, NY, USA
[5] Cornell Tech, New York, NY, USA
[6] Institute for Intelligent Biotechnologies (iBIO), Helmholtz Center Munich, Neuherberg, Germany
[7] Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians University Munich, Munich, Germany

Abstract. Light sheet fluorescence microscopy (LSM) enables high-resolution, three-dimensional (3D) imaging of biological specimens, providing rich volumetric data for studying cellular organization, pathology, and vascular networks. However, the size, dimensionality, and annotation burden of LSM data make supervised deep learning approaches costly and difficult to scale. Additionally, despite the abundance of unannotated LSM volumes, foundation models for this modality remain underexplored due to computational challenges and the complexity of volumetric representation learning. In this work, we introduce a 3D foundation model for LSM data, pretrained on a large curated collection of 3D images spanning multiple organisms, stains, and imaging protocols. We learn transferable volumetric representations by jointly optimizing for masked reconstruction and image-text alignment. The pretrained backbone drastically reduces the annotation burden, enabling efficient, few-shot adaptation for varied downstream tasks. We evaluate this approach on downstream segmentation, classification, and deblurring. Our results demonstrate consistent improvements over baselines, (1) when measured using standard evaluation metrics and (2) when rigorously assessed by domain experts. This highlights the potential of foundation model pretraining to reduce annotation requirements while improving performance across diverse LSM analysis tasks. Pretrained model weights and code for pretraining and finetuning are publicly available: https://github.com/AdinaScheinfeld/lsm_fm_public_repo.git.

Keywords: Foundation Models · LSM · SSL · Multimodal Learning.

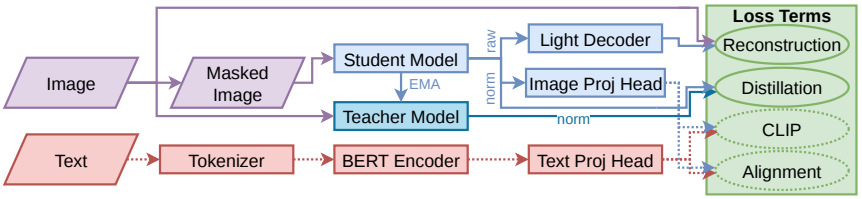

Fig. 1. Overview of pretraining framework. A student-teacher architecture (UNet or SwinUNETR) is trained with masked image reconstruction, exponential moving average (EMA)-based distillation, where teacher weights are updated as a moving average of the student, and CLIP-style image-text alignment using BERT-encoded text embeddings [24]. Dashed lines are associated with the text branch which can be removed for image-only model pretraining.

1 Introduction

Light sheet fluorescence microscopy has become a valuable imaging tool, enabling the rapid acquisition of high-resolution 3D images at whole-organ [29,14,35] and whole-organism scale [19]. The size and quality of these images enable analysis of cellular structures, pathological markers, and complex networks, yet also present challenges for downstream analysis due to the time and expertise required. And while deep learning methods have shown promise in accelerating such analyses, their success typically depends on large quantities of annotated training data [18,23]. For LSM, generating annotations is difficult as segmentation requires voxel-level labels, image classification requires expert knowledge of stains and structures, and image restoration tasks require paired degraded and high-quality volumes. Therefore, most existing models are trained for narrow tasks (e.g. segmentation only [1,4,23]) or for specific structures (e.g. nuclei or vessels only [1,23,31]) and fail to generalize across diverse staining protocols and biological contexts.

Foundation models offer a compelling alternative by learning transferable representations from large-scale unlabeled data through self-supervised objectives [5]. Such models have achieved success in domains including natural images [16,33], medical imaging [17,36,38], genomics [7], language [9], and agentic clinical reasoning [30]. In microscopy however, foundation model development remains limited, particularly for volumetric light sheet data. Owing to its high signal-to-noise ratio and the abundance of unannotated volumetric data routinely generated in biological studies, LSM presents unique, underexplored opportunities for unsupervised learning.

Prior 2D approaches fail to leverage volumetric context, often producing jagged artifacts along the axial dimension [23,34]. To address these limitations, we introduce a *3D foundation model* for LSM that leverages unannotated images and free-text descriptions to learn rich multimodal representations. Our pretraining combining image modeling and reconstruction with image-text alignment allows the model to capture both structural and semantic information. The

resulting pretrained backbone can be finetuned with minimal annotated data to perform segmentation, classification, and deblurring. An overview of the pretraining strategy is presented in Fig. 1. Our contributions are threefold:

1. We curate a heterogeneous dataset of unannotated 3D LSM volumes across multiple organisms, stains, and imaging protocols, and compose corresponding textual descriptions for each sample.
2. We utilize our assembled vision-language dataset to pretrain an LSM foundation model that can be finetuned for various downstream tasks.
3. We evaluate image+text and image-only pretraining across downstream tasks and show that pretrained models outperform trained-from-scratch models and baselines.

2 Method

2.1 Pretraining Data

Datasets To promote generalization across organisms, structures, and imaging protocols, we assembled a heterogeneous pretraining dataset from multiple internal and publicly available resources.

The internal dataset contains 24 single-channel mouse whole-brain images, each with a distinct immunostain. From each image, ten representative patches were manually selected. These training patches will be made public.

Public datasets include the SELMA3D MICCAI challenge dataset [26,27] and three datasets from the Allen Institute [2,3,20]. The SELMA3D dataset includes 30 single-channel mouse and human brain images of diverse structures, including neurons, cell nuclei, amyloid plaques, chondrocytes, and astrocytes, and 9 dual-channel vessel images. The datasets from the Allen Institute, a human brain dataset [20], a developing mouse brain dataset [2], and a viral labeling dataset of projection neurons [3], each include numerous multi-channel images. From each public dataset, up to 10 non-overlapping patches were randomly selected per channel from each image, ensuring a minimum foreground threshold of 5%. In total, 1,023 volumetric 96^3 voxel patches were collected.

Augmentations All pretraining images were normalized using percentile-based intensity rescaling to map intensities to the [0, 1] range. To increase robustness and variability, extensive data augmentation was applied to each patch, including random flips, rotations, affine transformations, Gaussian noise, Gaussian smoothing, intensity scaling, and intensity shifting. Intensities were clamped back to [0, 1] after augmentation. Examples are shown in Fig. 2.

Text Prompts For each volume, domain experts composed a structured 2-4 sentence free text caption describing biologically and visually salient features. Captions describe the staining target (e.g., protein or cell type), imaging modality and channel, specimen source (organism and tissue), morphological characteristics (e.g., filamentous, punctate, tubular), spatial organization within the tissue, and relevant pathological context. This process ensured that descriptions reflected both domain-specific terminology and observable image features.

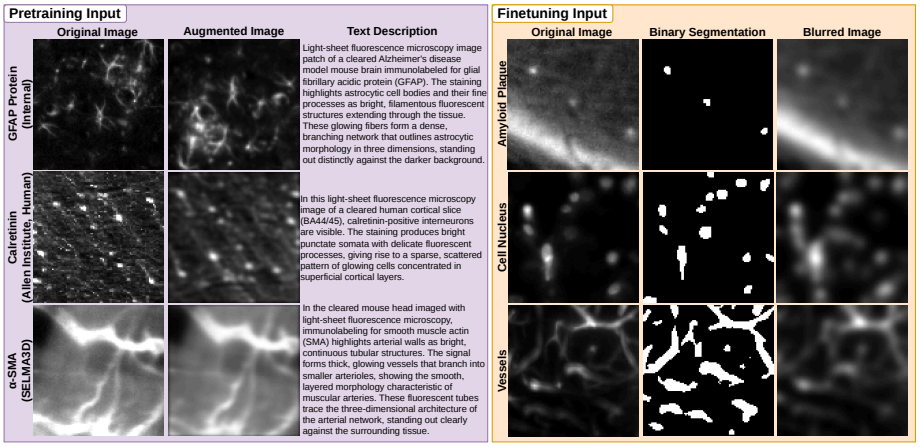

Fig. 2. Examples of input data used during pretraining (left), and finetuning (right). During pretraining, each volumetric patch is paired with a text description, forming image-text pairs for contrastive learning. During finetuning, the pretrained encoder is adapted for downstream tasks using paired images and corresponding labels, including binary masks for segmentation and synthetically blurred images for deblurring.

To increase linguistic diversity while preserving semantic fidelity, captions were paraphrased using an LLM [6]. Examples are shown in Fig. 2.

2.2 Finetuning Data

Datasets For downstream segmentation and deblurring experiments, three annotated datasets from the SELMA3D challenge were used [26,27]. Datatypes include amyloid plaque (single-channel), cell nucleus (single-channel), and vessel (dual-channel). Per structure and channel, up to 25 patches of size 96^3 with sufficient foreground representation were randomly sampled, yielding a total of 84 annotated patches. Examples are shown in Fig. 2.

The finetuning datasets were selected to represent distinct structural regimes common in LSM, allowing us to evaluate the generality of the learned representations. Amyloid plaques and cell nuclei both correspond to isolated foreground objects but differ in foreground density and spatial distribution, with amyloid plaques being sparse, punctate structures as compared with the dense foreground distribution of cell nuclei. Vessel images were included to represent a complementary domain characterized by contiguous, elongated structures. Unlike isolated objects, vascular structures require modeling long-range spatial continuity and topology, posing distinct challenges [28].

To increase the number of stain classes for classification experiments, images from three additional datasets were included, in addition to those used for segmentation and deblurring: c-Fos positive active neurons from SELMA3D [26], mouse brain images from the mesoSPIM initiative [32], and human brain images

from the Allen Institute [15]. In images with multiple channels, each stain was treated independently, and patches from all images were randomly selected ensuring sufficient foreground, resulting in 12 categories of images for classification.
Augmentations The intensity values in finetuning images were normalized identically to pretraining, and moderate augmentations including random flips, rotations, affine transformations, and Gaussian noise, were applied as well.

2.3 Model Architecture

Model Backbones We develop our foundation models using two 3D architecture families: a convolutional encoder-decoder model, UNet [25], and a hierarchical transformer-based model, SwinUNETR [10,21]. We instantiate the backbone twice for each architecture in order to construct a student-teacher framework. During training, we provide the student with masked 96^3 input volumes, where a randomized subset of non-overlapping patches of size 4^3 or 8^3 was zeroed out, while the teacher processes the corresponding unmasked volume. Rather than updating the teacher through gradient descent, we maintain its weights as an exponential moving average (EMA) of the student parameters. This asymmetric design encourages the student to infer stable representations at masked locations, while mitigating representation collapse [12,22].
Pretraining Objective The pretraining objective combines four complementary loss terms. (1) A knowledge distillation loss computes the KL divergence between the student and teacher feature distributions at masked positions, encouraging the student to recover the teacher's full-context representations from partial observations [37]. (2) A lightweight convolutional decoder, supervised by an L1 reconstruction loss on masked regions, maps the student's raw feature map back to voxel space. (3) A cosine alignment loss minimizes the angular distance between image embeddings and text embeddings produced using a pretrained BERT-based text encoder [8]. (4) A CLIP-style symmetric cross-entropy contrastive loss pushes matching image-text pairs together and non-matching pairs apart, across the batch [24]. The framework supports image-only pretraining by disabling the text stream, whereby the alignment and contrastive losses are zeroed and only the distillation and reconstruction objectives remain active.

The total loss is computed as a weighted sum of all four loss terms, with independently tunable coefficients for each component, as presented in Eq. (1). The combined gradient signal from the total loss is backpropagated through the student encoder, image projection head, light decoder, and (post-warmup) the unfrozen text encoder layers, while the teacher remains outside the computational graph.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{dist}}\,\mathcal{L}_{\text{dist}} + \lambda_{\text{rec}}\,\mathcal{L}_{\text{rec}} + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}} + \lambda_{\text{clip}}\,\mathcal{L}_{\text{clip}} \tag{1}$$

where $\gamma_{\text{dist}}, \gamma_{\text{rec}}, \gamma_{\text{align}}$, and γ_{clip} are independently tunable coefficients.

Training All trainable parameters are optimized jointly with AdamW. Training spanned 10-12 hours (280-500 epochs) on 2 A100/H100 GPUs using a 90/10

Table 1. Segmentation performance in a few-shot (number of train patches=5) and many-shot (number of train patches=15) settings. Values are averaged across two held-out test patches from each of three CV folds. Overall, pretrained models outperform trained-from-scratch counterparts and strong baselines.

		Amyloid Plaque				Cell Nucleus				Vessels			
		Few-shot		Many-shot		Few-shot		Many-shot		Few-shot		Many-shot	
	Method	Tot	Inst	Tot	Inst	Tot	Inst	Tot	Inst	Tot	Inst	Tot	Inst
Bases	Cellpose (2D) [23]	–	–	–	–	0.54	0.20	0.57	0.43	–	–	–	–
	Cellpose (3D) [23]	–	–	–	–	0.79	0.80	**0.82**	0.82	–	–	–	–
	CellSeg3D [1]	–	–	–	–	0.51	0.71	0.51	0.74	–	–	–	–
	uSAM (b) [4]	0.66	0.41	0.71	0.47	0.49	0.00	0.52	0.10	0.86	0.77	0.90	0.81
	uSAM (l) [4]	0.62	0.37	0.77	0.63	0.53	0.12	0.56	0.22	0.78	0.74	0.84	0.82
Scr	Swin	0.50	0.09	0.67	0.45	0.78	0.95	0.81	0.97	0.86	0.79	0.88	0.76
	UNet	0.50	0.11	0.72	0.58	0.75	0.86	0.80	0.88	0.83	**0.86**	0.89	0.68
Img	Swin	0.58	0.21	0.79	0.73	0.76	0.94	0.80	0.94	0.87	0.82	0.89	0.81
	UNet	0.50	0.03	0.50	0.00	**0.80**	0.87	**0.82**	0.96	0.61	0.56	0.89	0.84
Img+T	Swin	0.48	0.01	0.76	0.72	**0.80**	0.93	0.80	**0.99**	0.87	0.75	**0.92**	0.87
	Swin (over)	0.61	0.47	0.79	**0.83**	0.79	**0.96**	0.81	0.98	**0.89**	0.84	**0.92**	**0.88**
	UNet	**0.68**	**0.58**	**0.80**	0.69	0.78	0.91	0.81	0.94	0.86	0.76	**0.92**	0.81

Note: **Tot**: total Dice; **Inst**: instance Dice; **Bases**: baselines; **Scr**: trained from scratch; **Img**: image-only pretraining; **Img+T**: image+text pretraining; **over**: overfit variant. Bold indicates best per column.

train-validation split. Loss weights and other hyperparameter values were selected via a sweep. Details are available in our GitHub: https://github.com/AdinaScheinfeld/lsm_fm_public_repo.git.

3 Finetuning and Experiments

We consider three downstream tasks: Voxel-wise binary *segmentation*, patch-level stain-type *classification*, and *deblurring* of degraded patches. For each task, both the UNet and the SwinUNETR pretrained backbones were evaluated.

Segmentation For the UNet backbone, the pretrained encoder weights are mapped directly into a MONAI UNet. The encoder is frozen for an initial warmup period and subsequently finetuned at a reduced learning rate, while the final decoder block is trained at the full learning rate throughout. We instantiate a full SwinUNETR model, loading the pretrained encoder weights and freezing them during an initial warmup phase. The encoder is then unfrozen and finetuned at a lower learning rate than the convolutional decoder. UNet models were trained using Dice-Focal loss, while SwinUNETR models used Dice-Cross Entropy. Loss functions were chosen based on preliminary experiments to ensure stable optimization across architectures. For both, optimization was performed using AdamW, with early stopping based on validation Dice.

Classification The pretrained encoder is finetuned with a linear classification head appended to the bottleneck representation. The deepest encoder feature map is globally average-pooled to produce a fixed-size descriptor, which is then projected to class logits via a single linear layer. The encoder is frozen for an initial warmup period, then optimized at a reduced learning rate, while the classification head is trained at the full learning rate throughout. Cross-entropy loss

Table 2. Classification and deblurring performance in few-shot and many-shot settings. Few/many correspond to train patches=56/105 for classification and train patches=5/15 for deblurring. Values are averaged across held-out test patches from three CV folds. For deblurring, PSNR (not shown) exhibits similar trends as SSIM.

		Classification					Deblurring (SSIM)					
		Few			Many		Amyloid	Plaque	Cell Nucleus		Vessels	
	Method	ACC	MACRO F_1		ACC	MACRO F_1	Few	Many	Few	Many	Few	Many
Bases	PCA	0.33	0.28		0.47	0.40	–	–	–	–	–	–
	ResNet[11]	0.36	0.27		0.49	0.43	–	–	–	–	–	–
Scr	Swin	0.40	0.33		0.47	0.44	0.61	0.66	0.79	0.85	0.80	0.90
	UNet	0.49	0.46		0.61	0.57	0.63	0.67	0.81	0.87	0.88	0.90
Img	Swin	0.51	0.46		0.53	0.48	0.57	0.65	0.81	0.86	0.73	0.69
	UNet	**0.71**	**0.69**		0.62	0.58	0.64	**0.69**	**0.86**	**0.89**	**0.89**	**0.92**
Img+T	Swin	0.40	0.35		0.65	0.61	0.58	0.64	0.79	0.85	0.85	0.90
	UNet	0.61	0.57		**0.74**	**0.69**	**0.65**	**0.69**	0.80	**0.89**	0.78	0.90

Note: **Bases**: baselines; **Scr**: trained from scratch; **Img**: image-only pretraining; **Img+T**: image+text pretraining; **ACC**: accuracy. Bold indicates best per column.

with inverse-frequency class weighting is applied to account for class imbalance. Models were optimized using AdamW with early stopping (validation accuracy).

Deblurring For each datatype, we provide paired sharp and synthetically blurred image patches. The model is trained to reconstruct the sharp volume from the blurred input by predicting an additive residual that is summed with the blurred image and clamped to the range [0,1]. The training objective combines four terms: an L1 reconstruction loss, a structural similarity loss (SSIM), a 3D gradient-based edge consistency loss that penalizes discrepancies in spatial gradient magnitudes, and a high-frequency loss term to preserve fine structural details. Both backbones are instantiated as full encoder-decoder networks and initialized from the respective pretrained checkpoints. Optimization is performed using AdamW with early stopping based on validation total loss.

Experimental Design To analyze scaling behavior, we evaluated finetuning via 3-fold cross-validation across few-shot and many-shot regimes. Training sizes ranged from 5-15 patches per datatype for segmentation and deblurring, and 56-105 samples for classification. Across all tasks, data was randomly split 80/20 for training and validation, with 2 patches per datatype held out for testing.

Baselines Segmentation baselines include other pretrained models, μSAM (base and large) [4] and CellSeg3D [1], and a state-of-the-art fully supervised model, Cellpose-SAM (2D and 3D variants) [23]. Classification baselines include a non-deep-learning approach, PCA, and a standard deep-learning-based model, 3D ResNet-18 [11]. Additionally, a model of each backbone type was trained from scratch for each task as a directly comparable non-pretrained comparison.

Overtraining To assess whether pretraining epochs limited downstream performance, we conducted an overtraining experiment where the SwinUNETR model was trained beyond validation convergence (∼4,000 epochs), effectively overfitting [13]. The final checkpoint from this model was then finetuned as before.

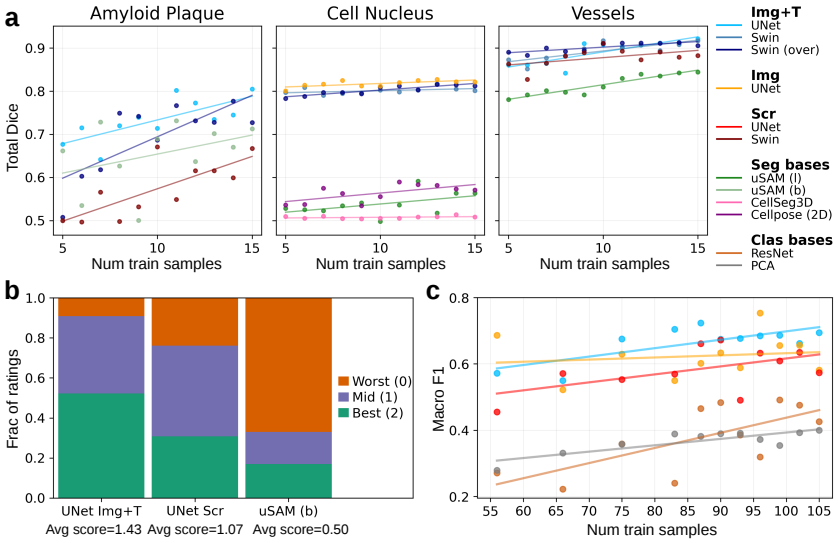

Fig. 3. Inference results (a) Segmentation performance (total Dice) across train sizes. Points show cross-validation averages on held-out patches. Pretrained models outperform scratch-trained models and baselines across datatypes. **(b)** Blinded expert evaluation of 60 segmentation samples. Bars show the distribution of best/middle/worst rankings (2/1/0 points), with average scores reported below each. The pretrained model achieves highest overall and average scores. **(c)** Classification performance (macro F_1) across train sizes, averaged over cross-validation splits. Points denote mean performance on held-out patches. Pretrained models outperform scratch-trained models and baselines. **Note**: **Img+T**: image+text pretraining; **Img**: image-only pretraining; **Scr**: trained from scratch; **uSAM (l/b)**: large/base uSAM models; **over**: overfit variant.

4 Results

Metrics-Based Evaluation Across all tasks and datasets, pretrained models consistently outperform trained-from-scratch counterparts and task-specific baselines, particularly in low-data regimes. Image-only pretraining already yields substantial gains, while image-text pretraining provides additional improvements in several settings, especially for segmentation.

Segmentation results from pretrained models show marked improvements in total Dice and instance Dice across amyloid plaque, nucleus, and vessel datasets as compared with trained-from-scratch models and baselines. Few-shot and many-shot results on held-out test sets are summarized in Table 1 and extended results are presented in Fig. 3. Similarly, classification performance improves in both accuracy and macro F1, with complementary strengths between image+text and image-only models, depending on train set size. Few-shot and many-shot results on held-out test sets are summarized in Table 2 and extended results are presented in Fig. 3. For deblurring, pretrained models achieve higher SSIM and

PSNR scores, indicating improved perceptual quality and fidelity. Results on held-out test sets are reported in Table 2.

Furthermore, across most downstream tasks, performance differences between the overtrained and normally trained backbones were modest, indicating that standard pretraining was sufficient. However, in select experiments, the overtrained backbone yielded small but consistent improvements.

Human Expert Evaluation Qualitative evaluation by 6 domain experts, each with 5-10 years of experience with LSM images, was conducted as an external validation of the quantitative findings. In blinded assessments, each expert independently evaluated 60 segmentation predictions spanning multiple datatypes. Across experts and datatypes, the pretrained model was most frequently selected as the best-performing approach, reinforcing the improvements observed in the metrics-based analysis and demonstrating that the quantitative gains correspond to perceptible improvements in segmentation quality, as highlighted in Fig. 3.

5 Discussion and Conclusion

We present a foundation model for light sheet microscopy that leverages unannotated volumetric data and text descriptions to enable efficient and improved few-shot and many-shot learning across downstream tasks. Our findings demonstrate that foundation model pretraining enables effective transfer to segmentation, classification, and deblurring tasks with minimal annotation. By reducing reliance on extensive annotations and improving generalization across stains and structures, our approach offers a practical pathway toward scalable and reusable analysis pipelines for large-scale microscopy data.

The strong performance of image-only pretrained models suggests that structural cues alone are highly informative, while text used in image+text models provides complementary semantic guidance that further improves results in specific scenarios. Moreover, while excessive pretraining is not strictly necessary, extended exposure to diverse unannotated volumes may further refine representations in certain settings, without degrading generalization.

Acknowledgement Research data generation from Zhuhao Wu laboratory in this report was supported by the NIH grant R01AI158676, R01MH131537, RF1MH128969 and R01MH142410.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achard, C., Kousi, T., Frey, M., Vidal, M., Paychère, Y., Hofmann, C., Iqbal, A., et al.: CellSeg3D: self-supervised 3D cell segmentation for microscopy. eLife **13**, RP99848 (2025)
2. Allen Institute for Brain Science: Developing Mouse Brain Imaging and Common Coordinate Framework,https://knowledge.brain-map.org/data/NUM7DTHI95ECV27X5RV

3. Allen Institute for Brain Science: Viral sparse labeling of connectionally-unique projection neurons for morphological assessment,https://knowledge.brain-map.org/data/VJGWTBZLG77YG5NKLKI
4. Archit, A., Freckmann, L., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., et al.: Segment Anything for Microscopy. Nature Methods **22**, 579–591 (2025)
5. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., et al.: On the opportunities and risks of foundation models. arXiv:2108.07258 (2022)
6. ChatGPT,https://chatgpt.com/
7. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods **21**, 1470–1480 (2024)
8. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 (2019)
9. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., et al.: The Llama 3 Herd of Models. arXiv 2407.21783 (2024)
10. Hatamizadeh, A., Nath, V., Tang, Y, Yang, D., Roth, H., Xu, D.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. arXiv:2201.01266 (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv 1512.03385 (2015)
12. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv:1503.02531 (2015)
13. Hong, L., Xie, S.M., Li, Z., Ma, T.: Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models. arXiv:2210.14199 (2022)
14. Kaltenecker, D., Al-Maskari, R., Negwer, M., Hoeher, L., Kofler, F., Zhao, S., Todorov, M., et al.: Virtual reality-empowered deep-learning analysis of brain cells. Nature Methods **21**(7), 1306–1315 (2024)
15. Kamentsky, L., Slayton, M., Park, J., Su-Arcaro, C., Moukheiber, M., Zhao, V.: Light sheet imaging of the human brain (Version draft) [Data set]. DANDI Archive (2023)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., et al.: Segment anything. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3992–4003 (2023)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024)
18. Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., et al. The multimodality cell segmentation challenge: toward universal solutions. Nature Methods **21**, 1103–1113 (2024)
19. Mai, H., Luo, J., Hoeher, L., Al-Maskari, R., Horvath, I., Chen, Y., Kofler, F., et al.: Whole-body cellular mapping in mouse using standard IgG antibodies. Nature Biotechnology **42**(4), 617–627 (2024)
20. Mazzamuto, G., Costantini, I., Gavryusev, V., Castelli, F.M., Pesce, L., Scardigli, M., Pavone, F.S., et al.: Human brain cell census for BA 44/45 (Version draft) [Data set]. DANDI archive (2022)
21. MONAI Documentation: Neural Network Architectures,https://monai-dev.readthedocs.io/en/fixes-sphinx/networks.html
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., et al.: DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193 (2024)

23. Pachitariu, M., Rariden, M., Stringer, C.: Cellpose-SAM: superhuman generalization for cellular segmentation. bioRxiv 2025.04.28.651001 (2025)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., et al.: Learning Transferable Visual Models for Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning, pp.8748–8763 (2021)
25. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 (2015)
26. SELMA3D 2024 Grand Challenge Dataset, https://selma3d.grand-challenge.org/
27. SELMA3D 2025 Grand Challenge Dataset, https://selma3d2025.grand-challenge.org/
28. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P.W., et al.: clDice - a Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. arXiv:2003.07311 (2022)
29. Stelzer, E.H.K., Strobl, F., Chang, BJ., Preusser, F., Preibisch, S., McDole, K., Fiolka, R.: Light sheet fluorescence microscopy. Nature Reviews Methods Primers **1**, 73 (2021)
30. Stoffl, L., Wiestler, B., Paetzold, J.C.: VERITAS: Verifiable Epistemic Reasoning for Image-Derived Hypothesis Testing via Agentic Systems. arXiv:2604.12144 (2026)
31. Todorov, M.I., Paetzold, J.C., Schoppe, O., Tetteh, G., Shit, S., Efremov, V., Todorov-Völgyi, K., et al.: Machine learning analysis of whole mouse brain vasculature. Nature Methods **17**(4), 442–449 (2020)
32. Voigt, F.F., Kirschenbaum, D., Platonova, E., Pagès, S., Campbell, R.A.A., Kastli, R., Schaettin, M., et al.: The mesoSPIM initiative: open-source light-sheet microscopes for imaging cleared tissue. Nature Methods **16**(11), 1105–1108 (2019)
33. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14408–14419 (2023)
34. Weigert, M., Schmidt, U., Haase, R., Sugawara, K., Myers, G.: Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3666–3673 (2020)
35. Zhao, S., Todorov, M.I., Cai, R., Al-Maskari, R., Steinke, H., Kemter, E., Mai, H., et al.: Cellular and Molecular Probing of Intact Human Organs. Cell **180**(4), 796–812 (2020)
36. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Kiblawi, S., Naumann, T., et al.: A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nature Methods **22**, 166–176 (2025)
37. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image BERT Pre-Training with Online Tokenizer. arXiv:2111.07832 (2022)
38. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**, 156–163 (2023)