Foundation Models

Preparing for the iPhone Moment in Medical Image Computing

MICCAI MedAGI, Marrakech, Morocco

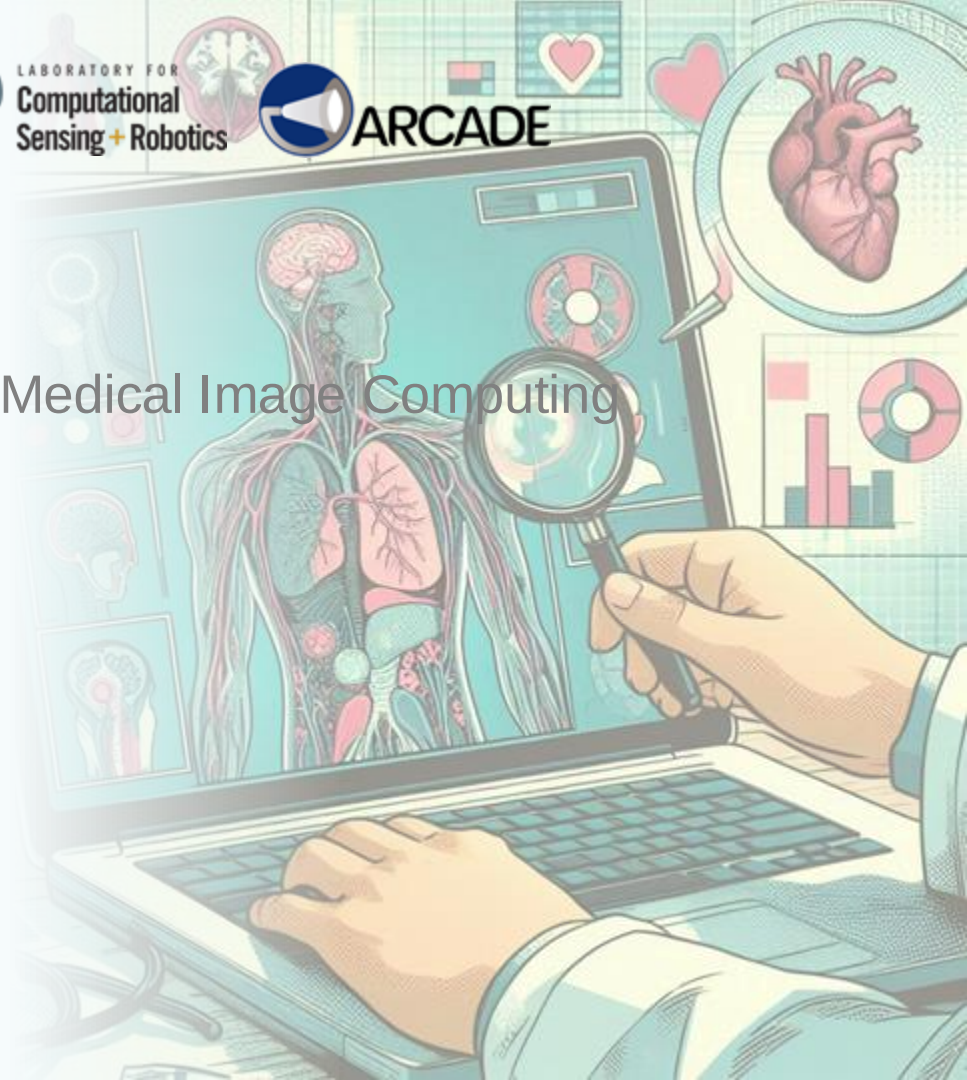
October 6th, 2024

Mathias Unberath, PhD

John C. Malone Associate Professor

Department of Computer Science

Johns Hopkins University



2007: The iPhone moment

→ New technology resulting in profound transformation

Thanks to Bing Image Creator running DallE-3

2022: The ChatGPT moment

→ New technology that transforms society!

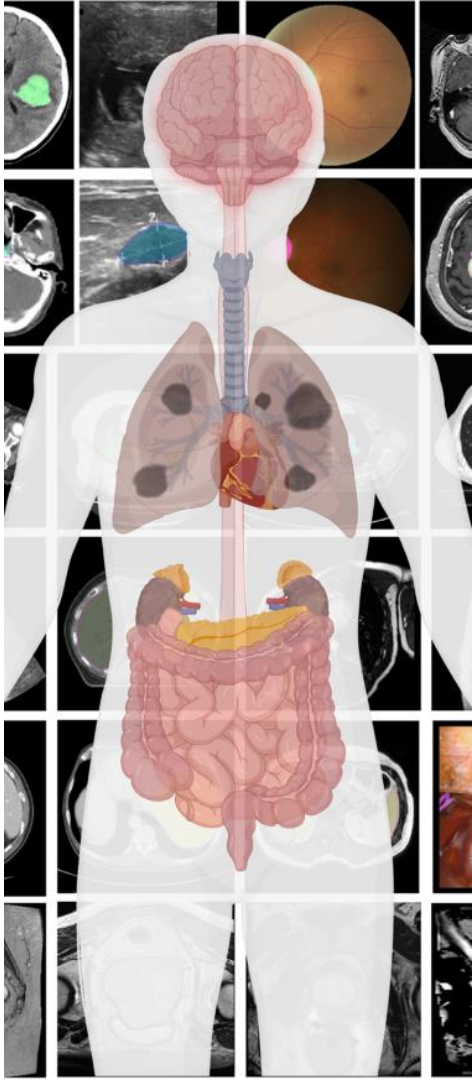
Foundation Models on the Rise

… and many more!

What makes foundation models different?

Formulated to solve “foundational tasks”

- No uber-specific task definition
- Building blocks rather than “complete” solutions



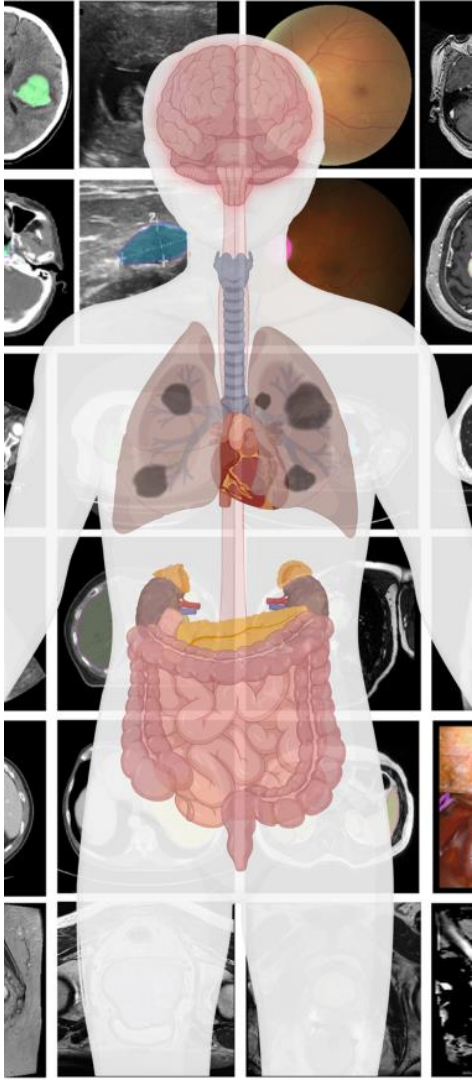
What makes foundation models different?

Formulated to solve “foundational tasks”

- No uber-specific task definition
- Building blocks rather than “complete” solutions

Flexibility enables scaling

- Dataset mixing (→ no “semantics”)
- Self-supervised learning



What makes foundation models different?

Formulated to solve “foundational tasks”

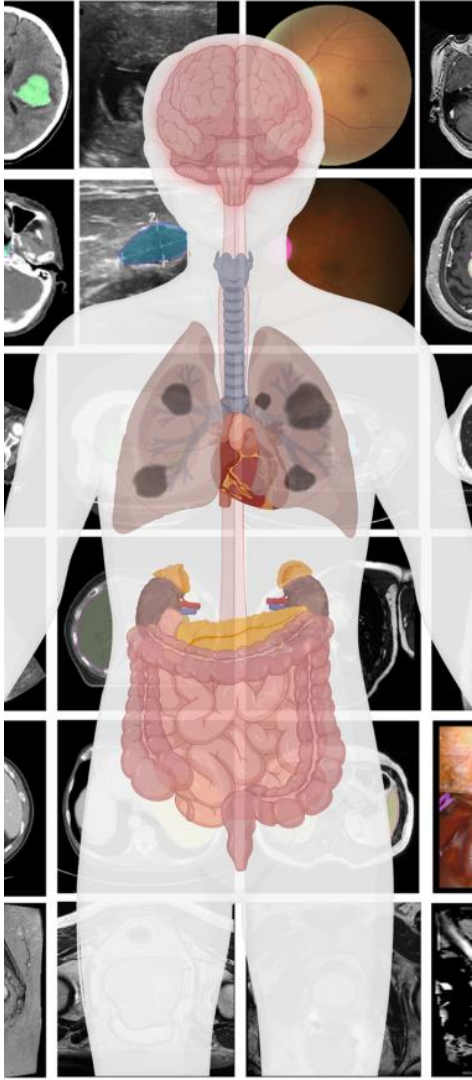
- No uber-specific task definition
- Building blocks rather than “complete” solutions

Flexibility enables scaling

- Dataset mixing (→ no “semantics”)
- Self-supervised learning

Scaling results in (better) generalization

- Larger variety in training sets
- Often: Applicable “out of the box” to new problem domains



Foundation Models – iPhone Moment for MIC?

Potential for catalytic effect in medical image computing and beyond

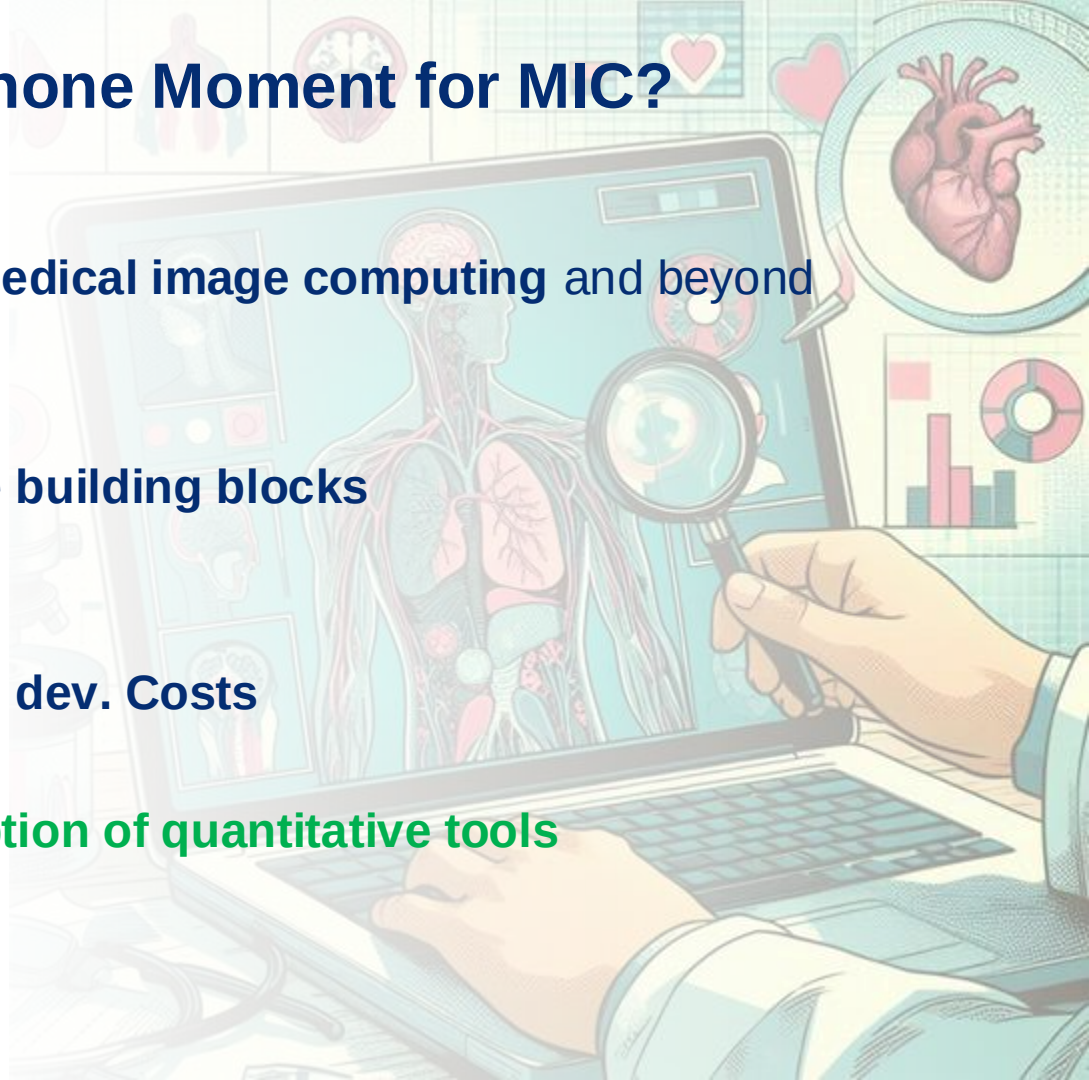
Flexible task formulations

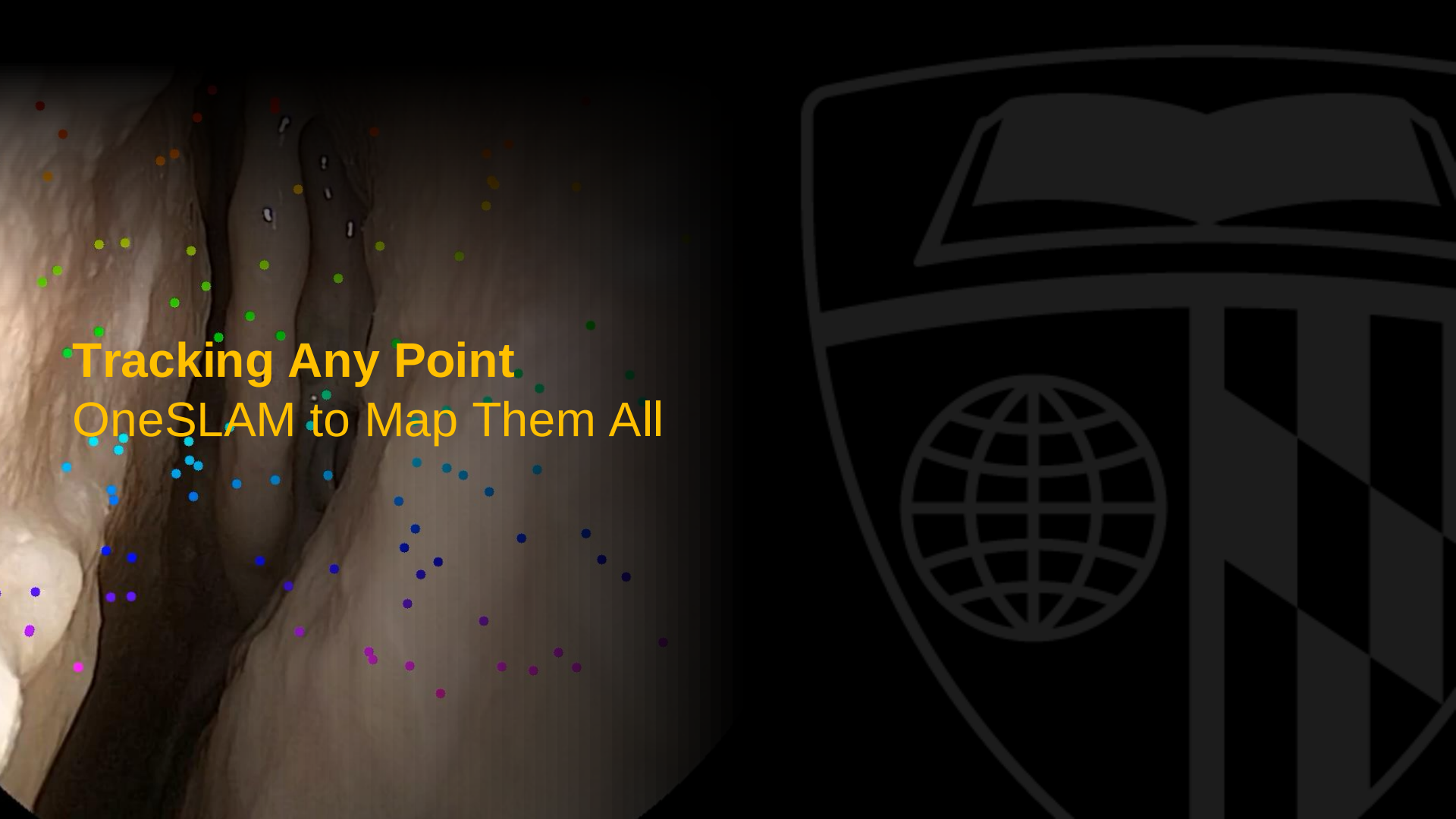
→ **Strong, generalizable building blocks**

“Out of the box generalization”

→ **Dramatically reduced dev. Costs**

→ **Potential for ubiquitous adoption of quantitative tools**




Tracking Any Point
OneSLAM to Map Them All

Quantitative Endoscopy

Endoscopy (and related modalities) is qualitative

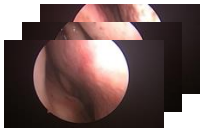
- Want: Endoscope tracking and reconstruction (SLAM)
- Enables quantitative analysis

Persistent challenges for computer vision in endoscopy

- Lack of photometric constancy (same looks different)
- Scarce features, repetitive texture

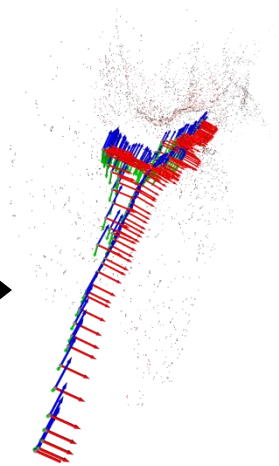
Liu, X., Stiber, M., Huang, J., Ishii, M., Hager, G. D., Taylor, R. H., & Unberath, M. (2020). Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment. In *MICCAI 2020*.

Opportunities for Foundation Models



Frame t

Machine Learning



Sparse reconstruction &
Pose estimates

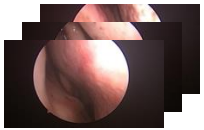
$P_t = [R_t|T_t]$

Current pose estimate

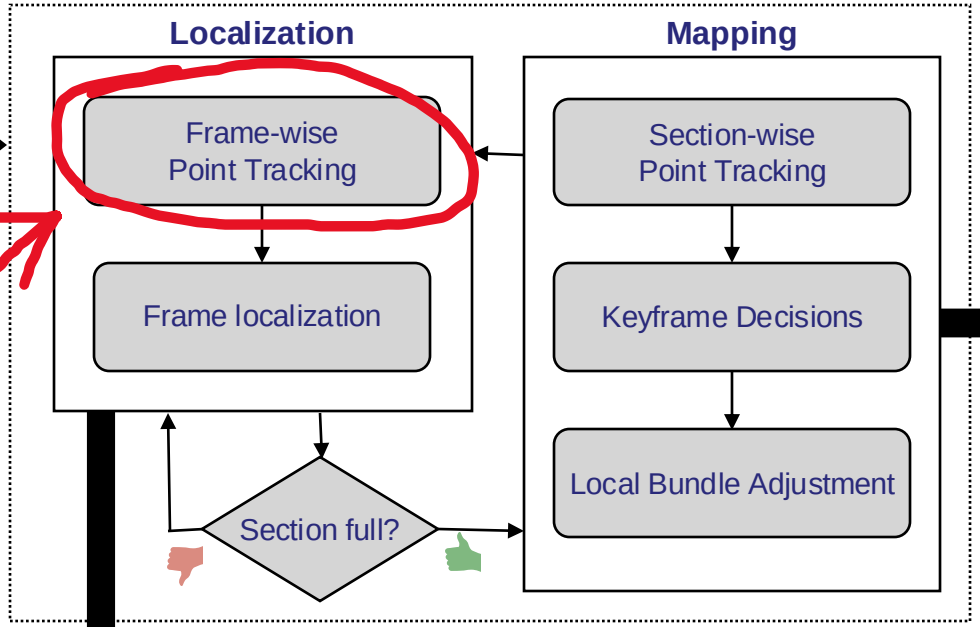
Highly specialized
Brittle solutions
→ Little flexibility

Liu, X., Zheng, Y., Killeen, B., Ishii, M., Hager, G. D., Taylor, R. H., & Unberath, M. (2020). Extremely dense point correspondences using a learned feature descriptor. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4847-4856).

Opportunities for Foundation Models

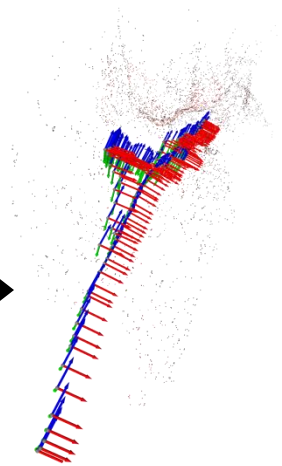


Frame t



$P_t = [R_t|T_t]$

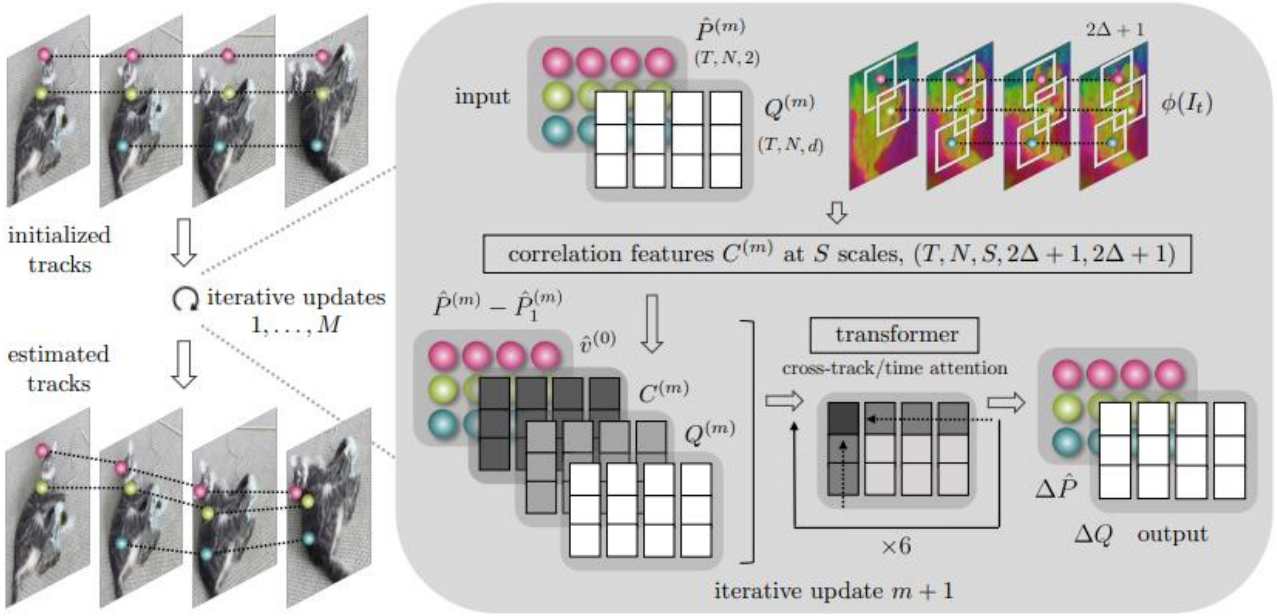
Current pose estimate



Sparse reconstruction &
Pose estimates

Liu, X., Zheng, Y., Killeen, B., Ishii, M., Hager, G. D., Taylor, R. H., & Unberath, M. (2020). Extremely dense point correspondences using a learned feature descriptor. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4847-4856).

Tracking Any Point Foundation Models

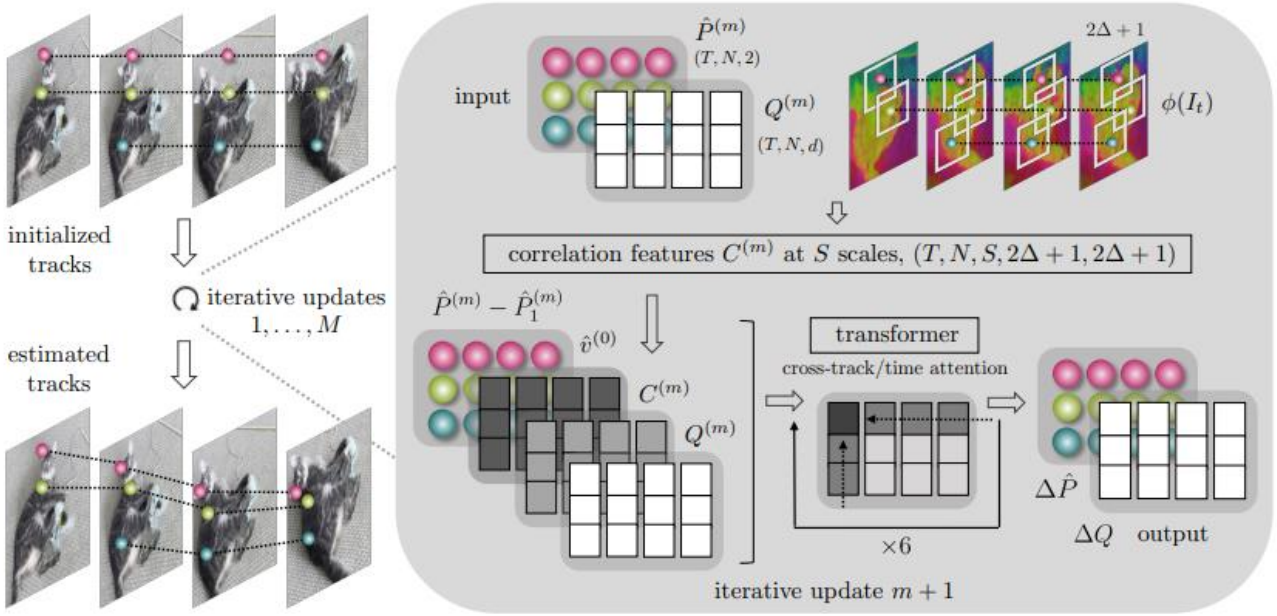


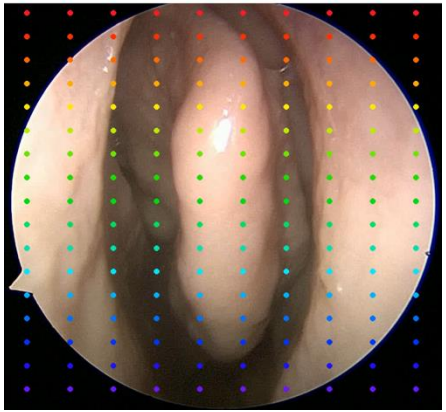
Iteratively update point tracks of arbitrary points

- Cross-track and cross-time attention
- Modern, transformer-based architecture

Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. arXiv:2307.07635 (2023)

Tracking Any Point Foundation Models



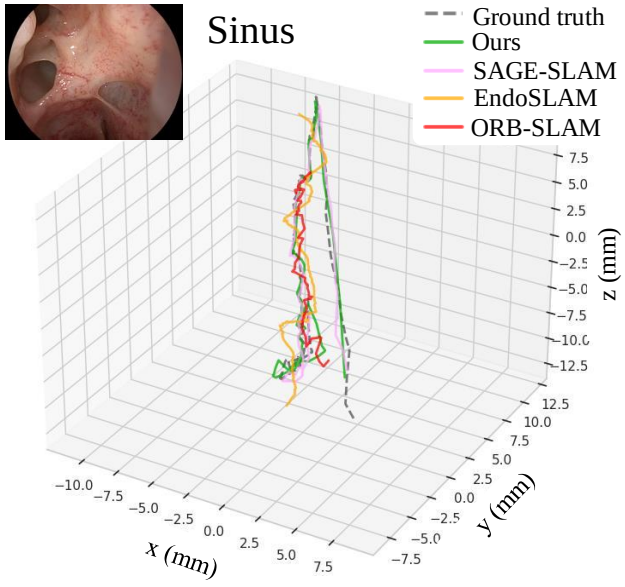


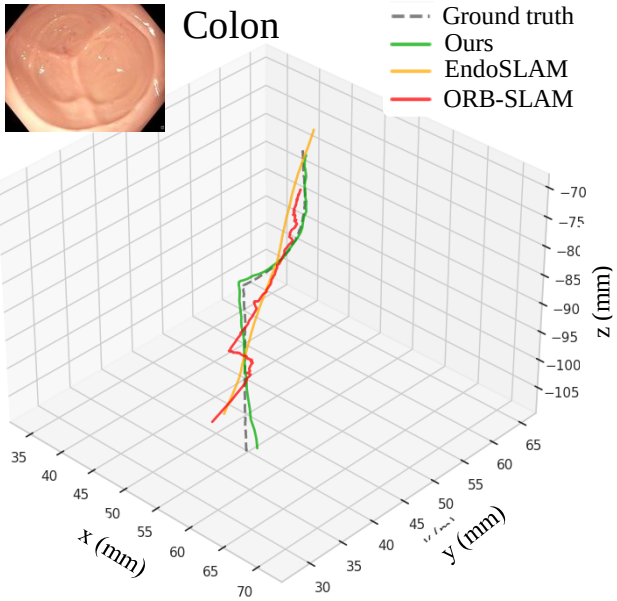
Iteratively update point tracks of arbitrary points

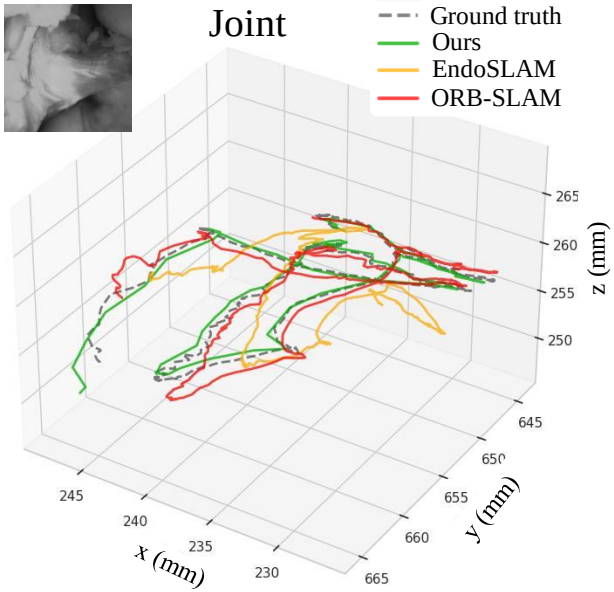
- Cross-track and cross-time attention
- Modern, transformer-based architecture

Seenivasan, L., …, Unberath, M.: Tracking Any Point Foundation Models for Endoscopic Applications, 2024. In Preparation.

SLAM using Tracking Any Point Foundation Models







OneSLAM based on Tracking Any Point Foundation Model

- First FM-based SLAM (at least for endoscopy)
- Outperforms both: Conventional approaches & domain-specific ML methods

Teufel, T., Shu, H.,... & Unberath, M. (2024). OneSLAM to map them all: a generalized approach to SLAM for monocular endoscopic imaging based on tracking any point. IPCAI, 1-8.

Segmenting Anything

FastSAM3D and ProtoSAM3D

Segment Anything Models (SAM)

Segmentation is a key enabling task

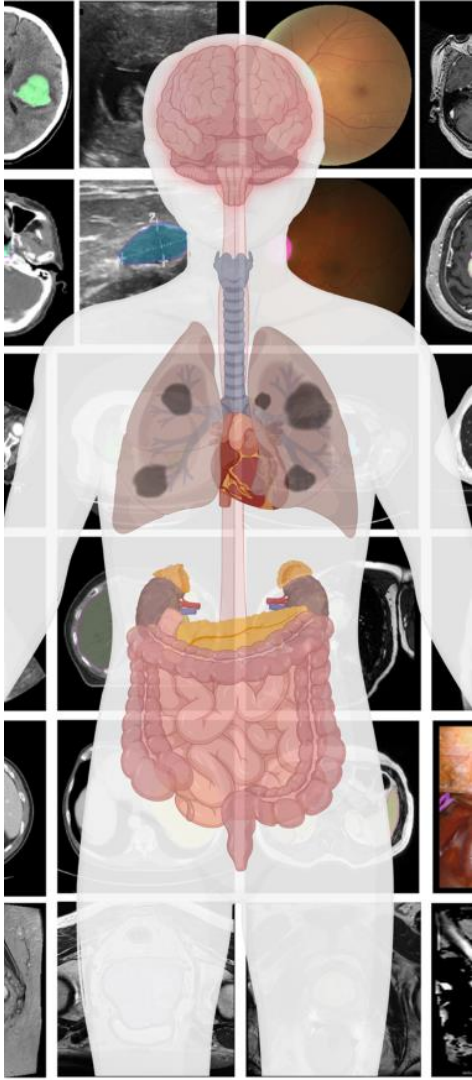
- “Making sense” of medical images
- Enables quantitative analysis

Foundation models for segmentation

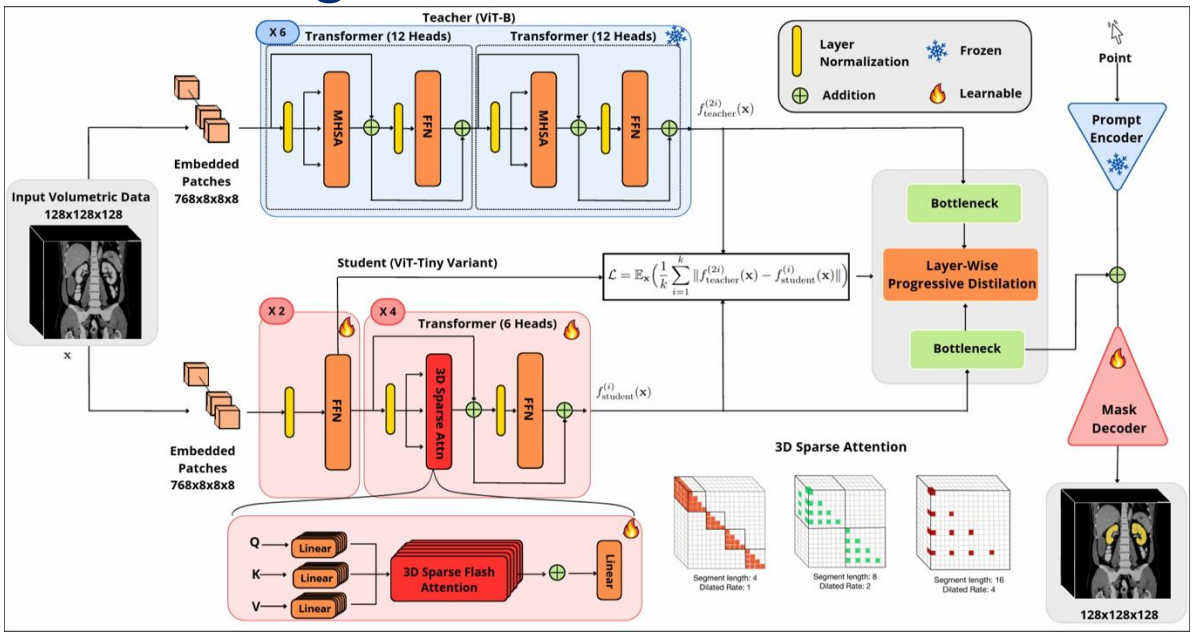
- Replace semantics with user query
- Interactivity ←→ Generalization

Two challenges

1. Scaling SAM to volumetric images
2. Clawing back semantic segmentation benefits



Scaling SAM to High-resolution Volumes



Iteratively update point tracks of arbitrary points

- Layer-wise progressive distillation to a VIT-Tiny
- 3D sparse flash attention

Shen, Y., Li, J., Shao, X., Romillo, B. I., Jindal, A., Dreizin, D., & Unberath, M. (2024). FastSAM3D: An Efficient Segment Anything Model for 3D Volumetric Medical Images. MICCAI.

FastSAM3D

Dim	Method	Resolution	Encoder			Decoder			Acceleration	
			Time (ms)↓	FLOPs (G)↓	Memory (Gb)↓	Time (ms)↓	FLOPs (G)↓	Memory (Gb)↓	To 2D ↑	To 3D ↑
2D	SAM [16]	1024 × 1024	3980	369.0	7.87	239	3.0	5.57	1.00×	/
	MobileSAM [32]	1024 × 1024	584	36.7	5.48	233	3.0	5.27	5.16×	/
	TinySAM [24]	1024 × 1024	609	36.7	5.48	246	3.0	5.27	4.93×	/
	MedSAM [18]	1024 × 1024	3983	369.0	7.87	241	2.9	5.57	1.00×	/
	SAM-Med2D [4]	256 × 256	1063	32.0	6.32	216	**0.21**	5.55	3.30×	/
3D	SAM-Med3D [27]	128 × 128 × 128	70	89.5	6.58	20	2.8	5.53	60.27×	1.00×
	FastSAM3D	128 × 128 × 128	**3**	**21.9**	**0.78**	**5**	2.8	**0.71**	**527.38×**	**8.75×**

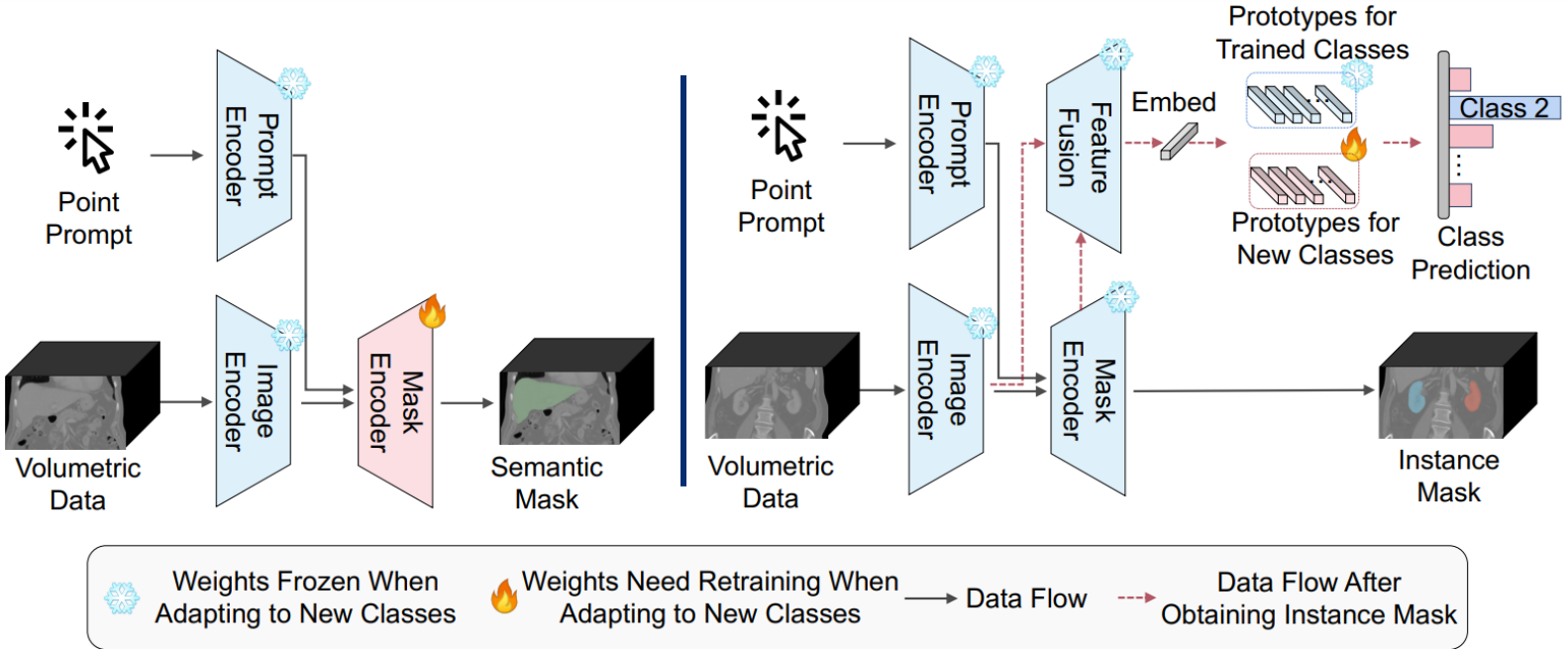
Dim	Method	AMOS [13]				TotalSegmentator [28]				BraTS [20]			
		1pt	3pt	5pt	10pt	1pt	3pt	5pt	10pt	1pt	3pt	5pt	10pt
2D	SAM [16]	0.049	0.093	0.114	0.145	0.202	0.279	0.311	0.348	0.108	0.192	0.217	0.237
	MobileSAM [32]	0.041	0.056	0.063	0.070	0.149	0.170	0.182	0.212	0.079	0.132	0.156	0.186
	TinySAM [24]	0.049	0.077	0.089	0.101	0.171	0.225	0.243	0.262	0.103	0.165	0.187	0.211
	MedSAM [18]	0.004	0.051	0.060	0.074	0.006	0.069	0.090	0.111	0.008	0.059	0.064	0.071
	SAM-Med2D [4]	0.097	0.127	0.129	0.132	0.008	0.081	0.100	0.128	0.013	0.076	0.082	0.084
3D	SAM-Med3D [27]	**0.289**	**0.386**	**0.418**	**0.448**	**0.252**	**0.400**	**0.463**	**0.522**	0.328	0.395	0.418	**0.446**
	FastSAM3D	0.273	0.368	0.402	0.437	0.250	0.378	0.445	0.519	**0.333**	**0.401**	**0.421**	0.445

Convincing performance (for SAM models)

- Best or second-best performance across several benchmarks
- 3ms execution! >500x / >8x speed-up compared to 2D / 3D methods

Shen, Y., Li, J., Shao, X., Romillo, B. I., Jindal, A., Dreizin, D., & Unberath, M. (2024). FastSAM3D: An Efficient Segment Anything Model for 3D Volumetric Medical Images. MICCAI.

Leveraging the Best of Both Worlds – ProtoSAM3D

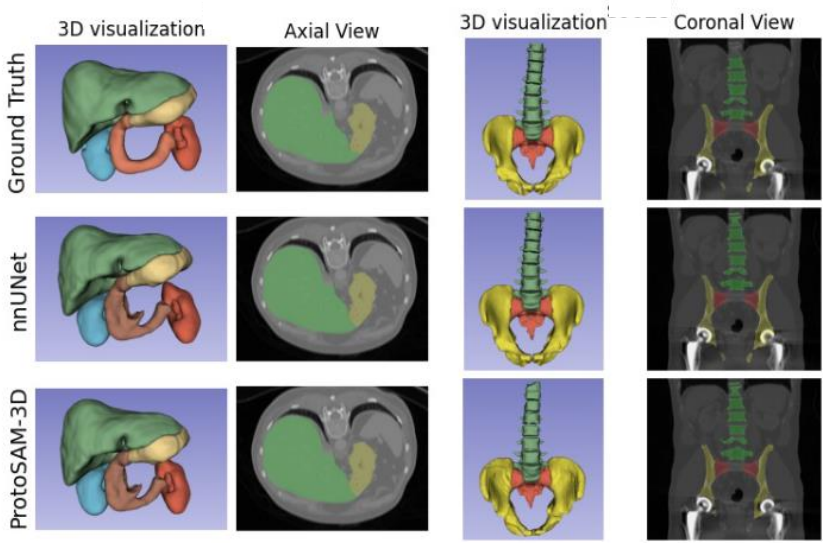


Automated semantic segmentation while retaining SAM's flexibility

- Two-stages: 1) Coarse prototype identification; 2) Fine segmentation
- Prompting still possible! New prototypes can be added anytime!

Shen, Y., Dreizin, D., Inigo Romillo, B., & Unberath, M. (2024). ProtoSAM3D: Interactive Semantic Segmentation in Volumetric Medical Imaging via a Segment Anything Model and Mask-Level Prototypes. Under review.

Leveraging the Best of Both Worlds – ProtoSAM3D



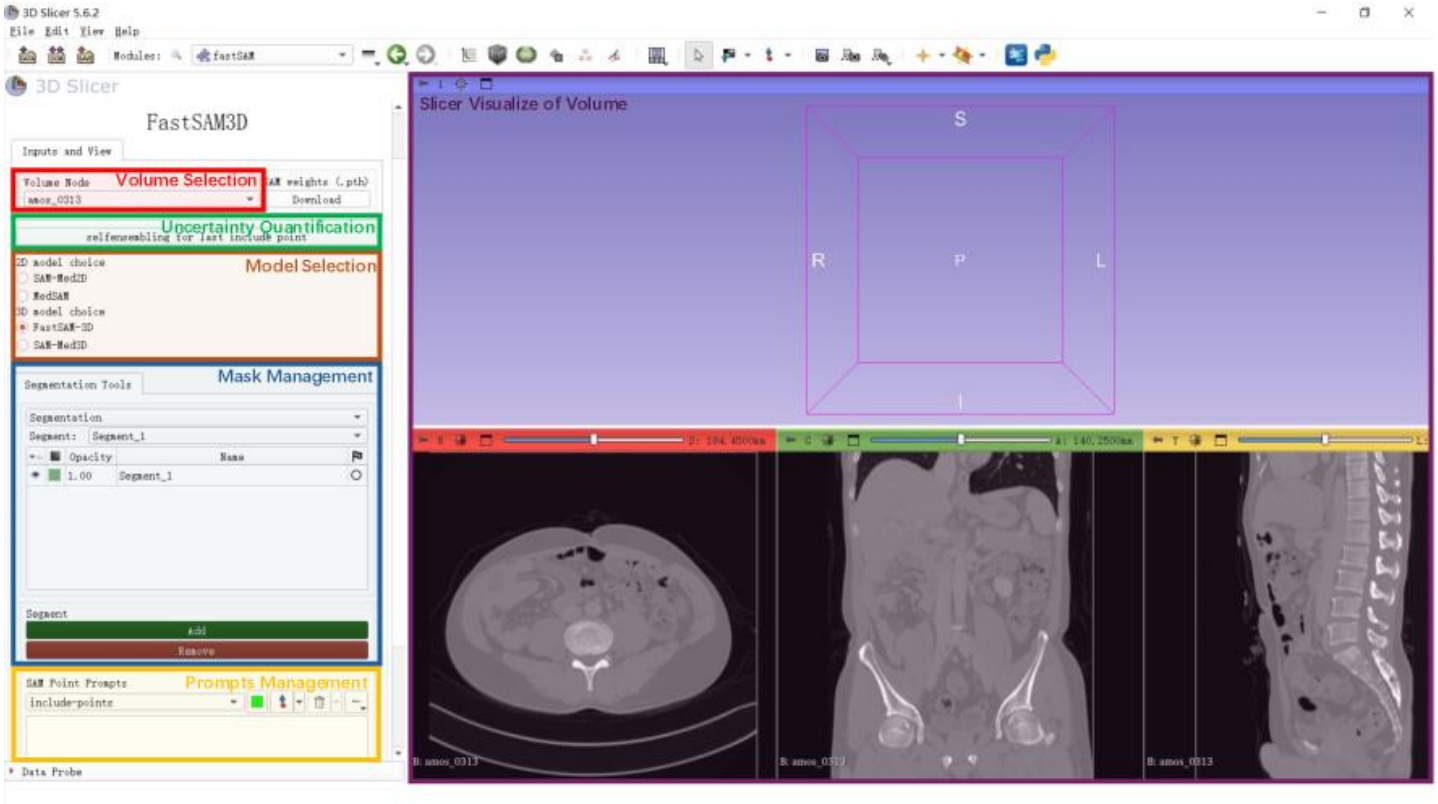
Methods	ΔParams. (M)	*AMOS-8* (Re) DSC	IoU	Acc	*AMOS-8* (Gen) DSC	IoU	Acc	*AMOS-7* (Zero) DSC	IoU
SwinUNETR-S (Hatamizadeh et al., 2021)	8.35	80.23	71.57	81.45	79.75	70.53	81.25	-	-
SwinUNETR-B (Hatamizadeh et al., 2021)	31.00	83.21	75.23	85.12	81.34	77.32	82.15	-	-
SwinUNETR-L (Hatamizadeh et al., 2021)	120.05	87.23	83.12	88.57	85.23	77.32	87.32	-	-
STU-Net-S (Huang et al., 2023b)	7.30	82.01	73.25	80.34	79.46	70.57	79.99	-	-
STU-Net-B (Huang et al., 2023b)	29.04	86.43	83.10	87.23	82.09	79.54	85.32	-	-
STU-Net-L (Huang et al., 2023b)	203.15	89.23	83.99	89.54	84.24	83.07	88.21	-	-
nnUNet (Isensee et al., 2021)	14.28	89.02	84.13	90.59	86.34	84.27	87.35	-	-
SAM (Kirillov et al., 2023)	0.00	-	-	-	35.21	26.07	-	36.43	23.76
MedSAM (Ma et al., 2024)	0.00	-	-	-	39.64	30.17	-	40.21	33.56
SAM-Med2D (Cheng et al., 2023)	0.00	-	-	-	40.12	34.74	-	41.35	33.07
SAM-Med3D (Wang et al., 2023)	0.00	-	-	-	45.42	36.52	-	44.76	31.56
SAM3D (Bui et al., 2024)	0.00	-	-	-	42.56	35.23	-	39.87	28.45
SAM-Med3D (F) (Wang et al., 2023)	7.58	85.24	78.54	86.76	84.23	78.32	86.34	-	-
Semantic-SAM (T) (Li et al., 2023a)	130.45	80.56	72.34	82.11	75.23	68.97	78.45	-	-
SAM-CLIP (T) (Wang et al., 2024)	198.32	78.23	70.12	80.56	73.45	66.78	76.89	-	-
ProtoSAM-3D w/o Distillation	0.03	89.78	84.78	92.12	87.34	85.21	89.32	45.76	**31.74**
ProtoSAM-3D w/ Distillation	**0.01**	**90.34**	**86.14**	**93.00**	**87.98**	**85.95**	**89.64**	**46.32**	**32.65**

Automated semantic segmentation while retaining SAM's flexibility

- Convincing qualitative and quantitative performance
- Among first models for semantic and zero-shot instance segmentation

Shen, Y., Dreizin, D., Inigo Romillo, B., & Unberath, M. (2024). ProtoSAM3D: Interactive Semantic Segmentation in Volumetric Medical Imaging via a Segment Anything Model and Mask-Level Prototypes. Under review.

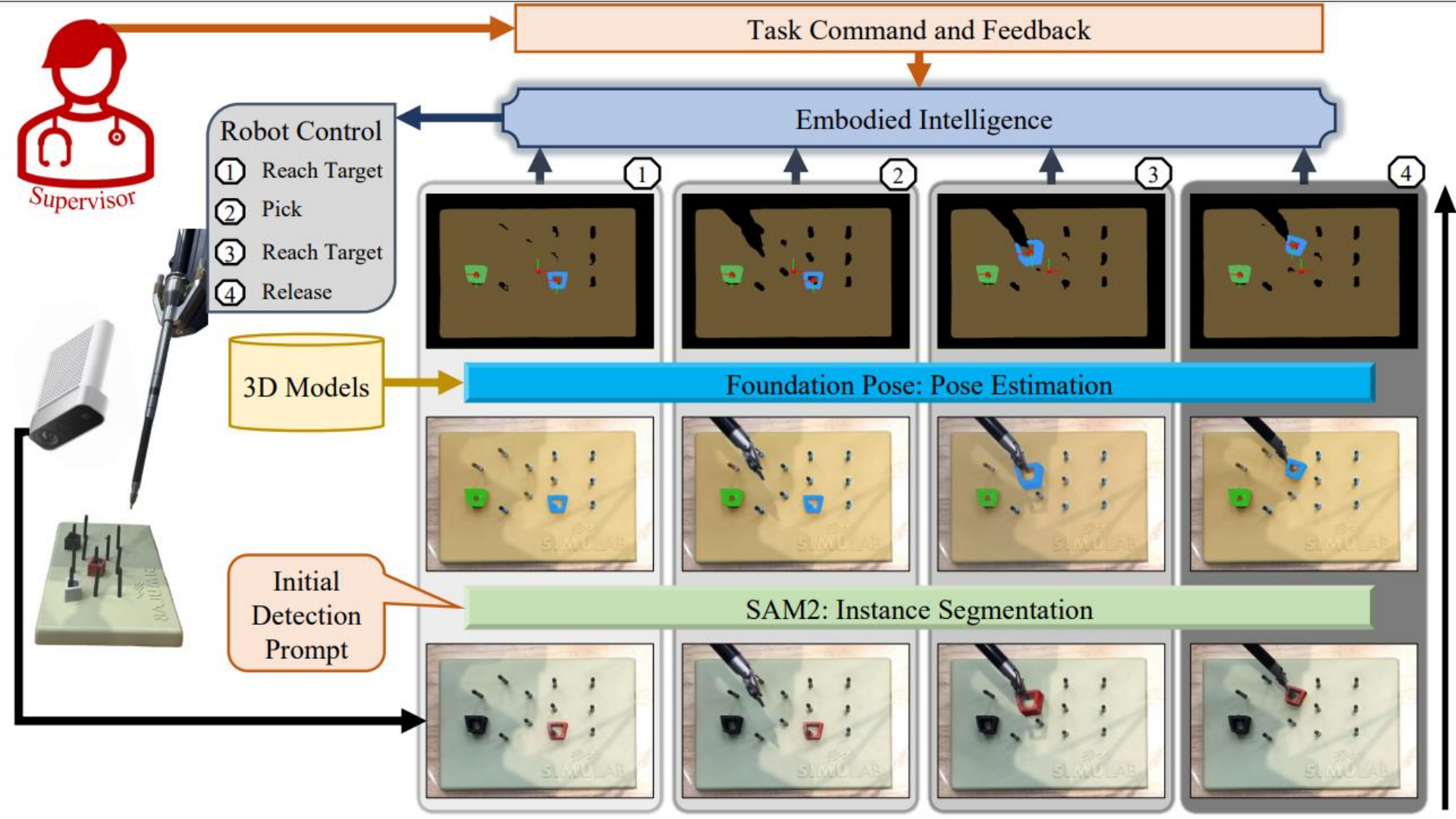
3DSlicer Interface for 2D and 3D SAM



Shen, Y., Shao, X., Romillo, B. I., Dreizin, D., & Unberath, M. (2024). FastSAM-3DSlicer: A 3D-Slicer Extension for 3D Volumetric Segment Anything Model with Uncertainty Quantification. *MICCAI MedAGI Workshop*.

Robot Automation

Catalytic Potential of Foundation Models as Building Blocks


Task Command and Feedback
Supervisor
Embodied Intelligence
Robot Control
1 Reach Target
2 Pick
3 Reach Target
4 Release
3D Models
Foundation Pose: Pose Estimation
Initial Detection Prompt
SAM2: Instance Segmentation
1
2
3
4

Towards Robust Automation of Surgical Systems via Digital Twin-based Scene Representations from Foundation Models

Hao Ding et al.

Department of Computer Science,
Johns Hopkins University

Concluding Thoughts

Towards an iPhone Moment in Medical Image Computing

Foundation Models for Medical Image Computing

This is the beginning – not the end!

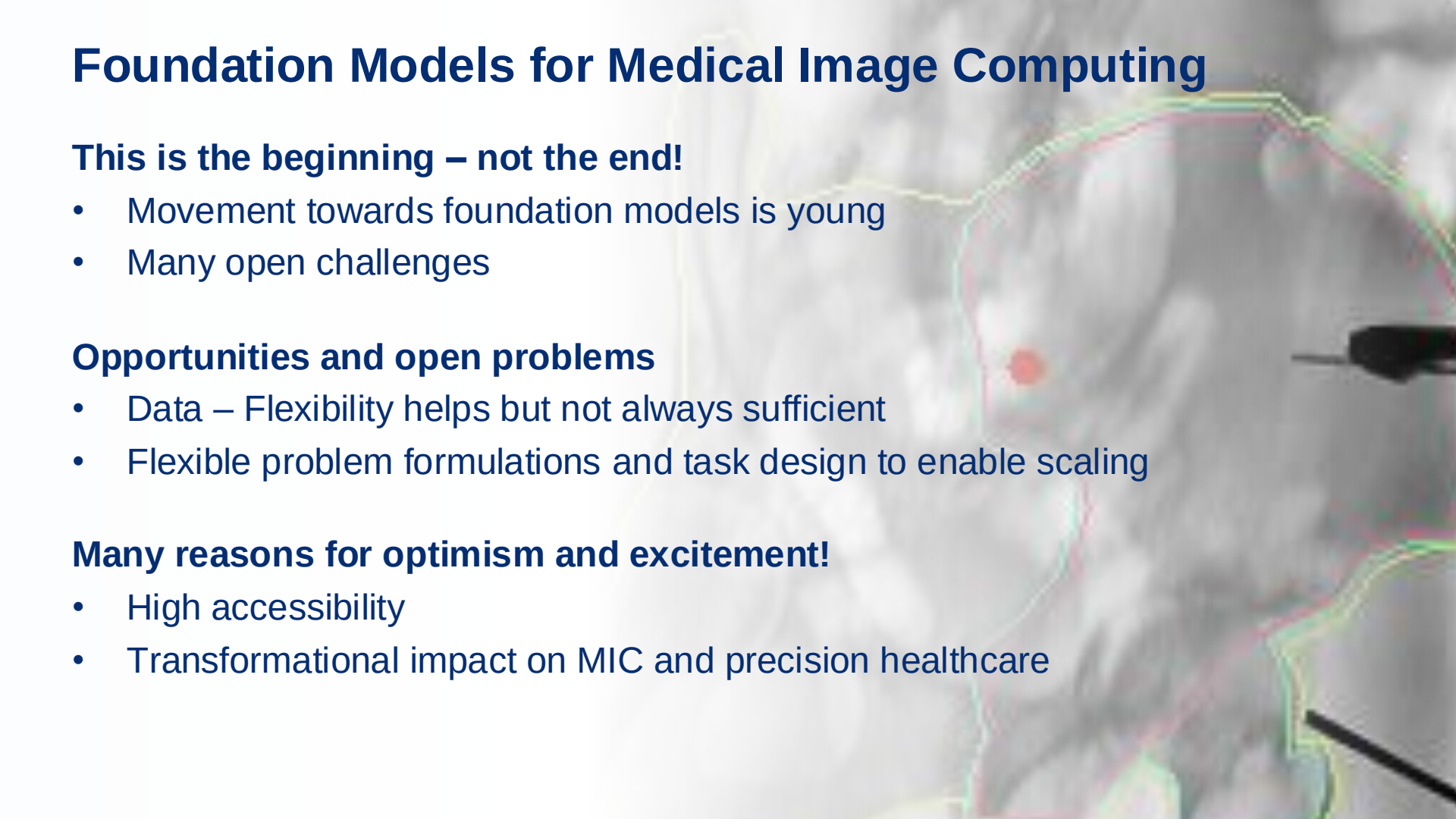
- Movement towards foundation models is young
- Many open challenges

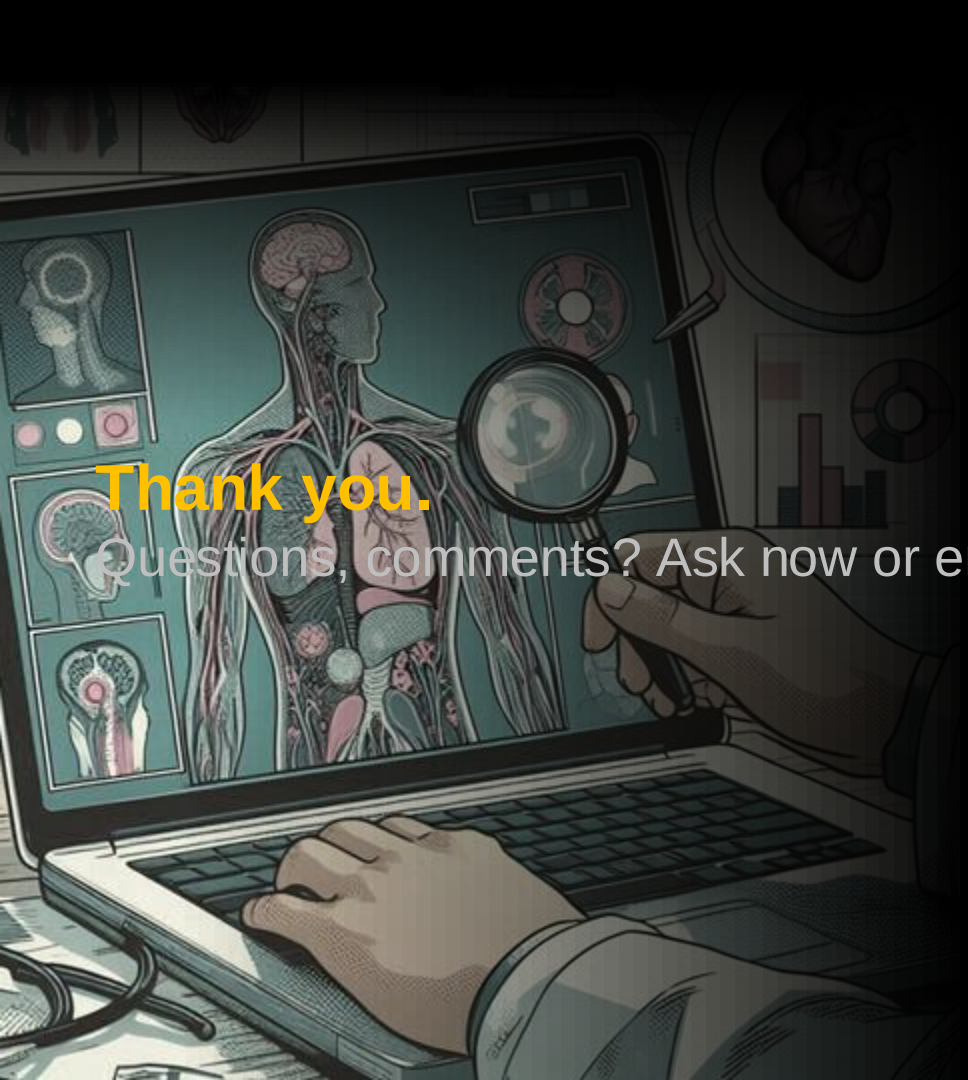
Opportunities and open problems

- Data – Flexibility helps but not always sufficient
- Flexible problem formulations and task design to enable scaling

Many reasons for optimism and excitement!

- High accessibility
- Transformational impact on MIC and precision healthcare





Thank you.

Questions, comments? Ask now or email me at: mathias@jhu.edu