RIKEN
Center for
Computational Science

Not all AI is GenAI; not all GenAI is LLMs

Case Studies of the Challenges in Using AI with Large Scale Scientific Computing

Mohamed Wahib

RIKEN Center for Computational Science, Japan

28th August 2025

Work of Many Collaborators

JP	RIKEN-CCS	Satoshi Matsuoka Kento Sato Jens Domke Jun Igarashi Aleksandr Drozd Emil Vatai Lingqi Zhang Peng Chen Sara Moukir
JP	Hokkaido U.	Masaharu Munetomo, Enzhi Zhang
JP	Science Tokyo	Toshio Endo Rio Yokota Riichiro Hira Takayuki Nishiyo Chen Zhuang Du Wu Tengfei Wang Ivan Ivanov
JP	AIST	Liu Xin , Truong Thao Nguyen
JP	Spring-8	Kentaro Uesugi, Takaki Hatsui
JP	Kyoto U.	Seo Akira, Yosuke Higo
SE	KTH	Artur Podobas
FR	DDN	Emmanuel Jeannot
	BSC	Jesus Labarta, Marc Clasca, Mario Acosta, Kai Keller
US/TR	Nvidia/Koç U.	Didem Unat, Ilyas Turimbetov
FR	Télécom SudParis	François Trahay
US	ANL	A. Dubey, I. Cloet, X. Wu, J. O'Neal, Klaus
US	UC Berkeley	Saathvik Selvan, Connor Chen
US	VirginiaTech	Johann Rudi, Wuchun Feng
US	ORNL	Xiao Wang, Issac Lyngaas
US	Emory U.	Katherine Keegan
US	Intel	Balazs Gerofi
HK	HKBU	Ameilie Zhou, Yuxin Wang
CN	NUIST	Xue Yu
SG	A*STAR	Luo Tao

Why at all is AI Relevant to Science/Engineering

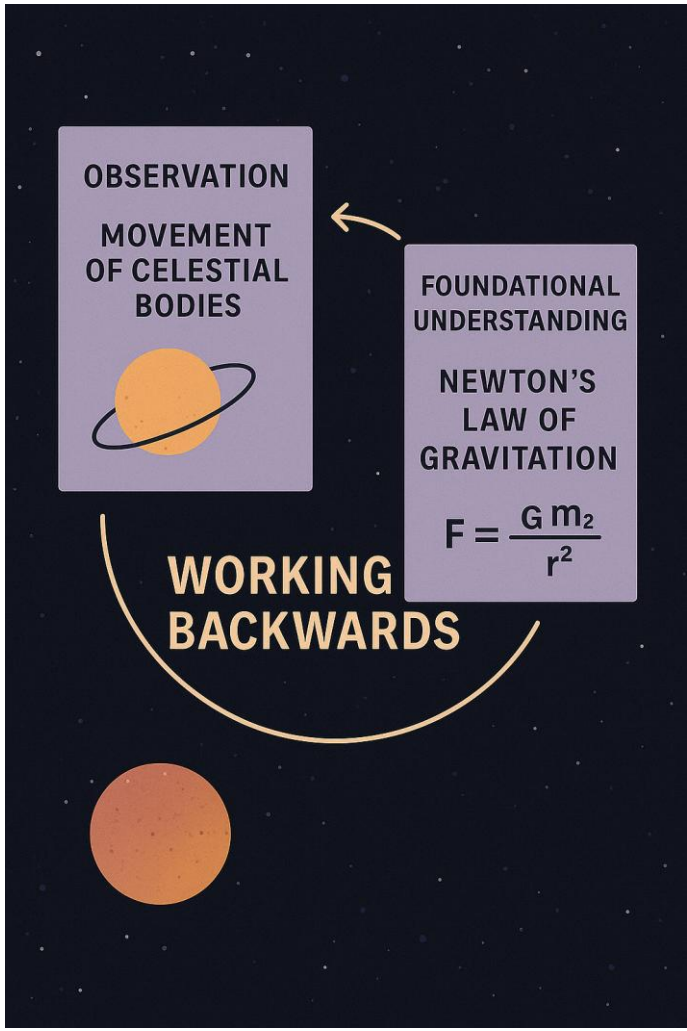
Astronomers observed **regular, predictable motion** of celestial bodies

↓

What governs this motion?

↓

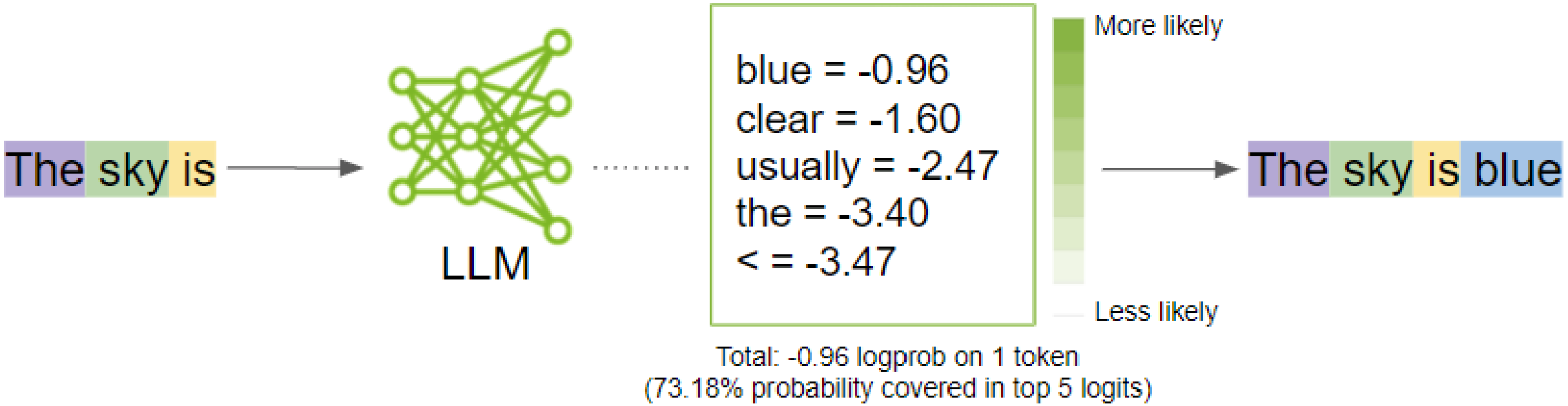
Newton worked backward from these observations and proposed: A **universal force** acts between all masses



Science is reverse engineering nature → (mostly) by observing patterns

AI (Machine Learning) is Good at Finding Patterns

When you use ChatGPT, you are asking the model to **predict from patterns** it learned

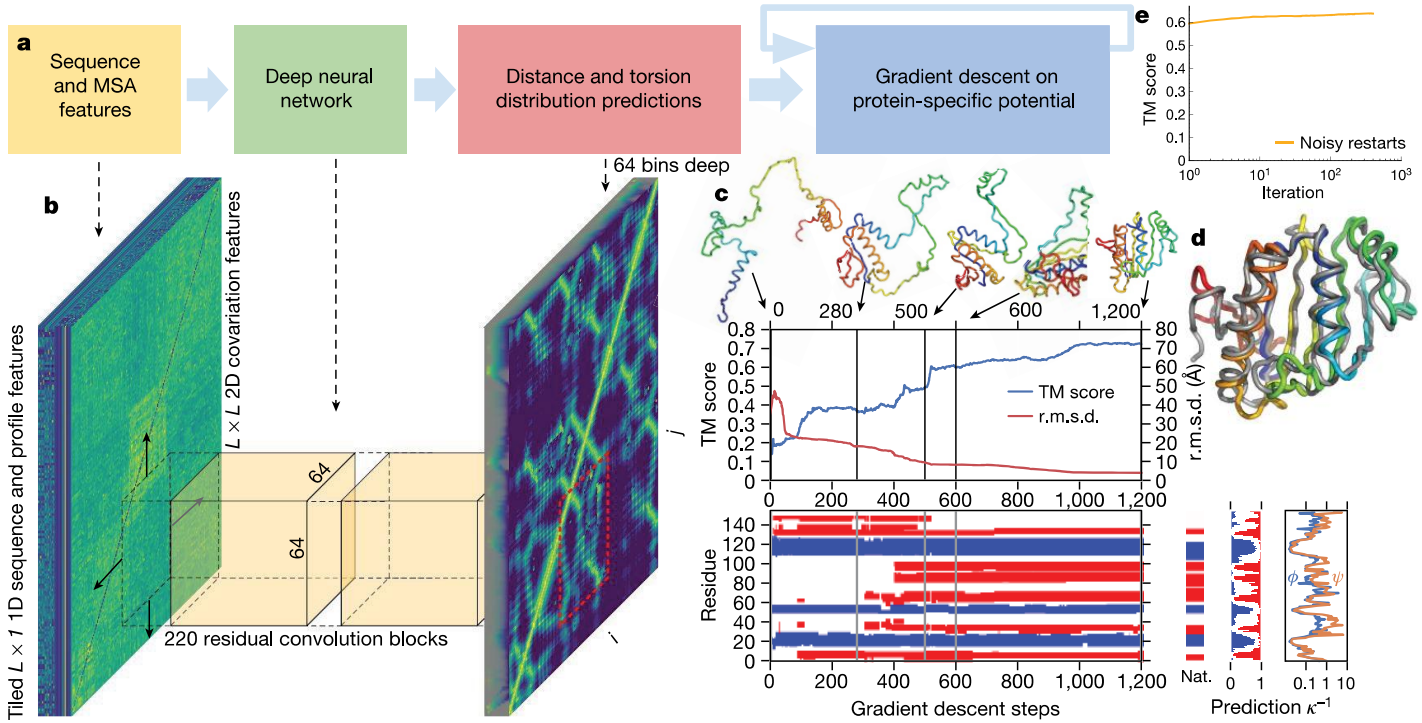


AI in Science: use AI pattern recognition capability to understand nature

Breakthroughs in AI-based Science: Finding Patterns

nature

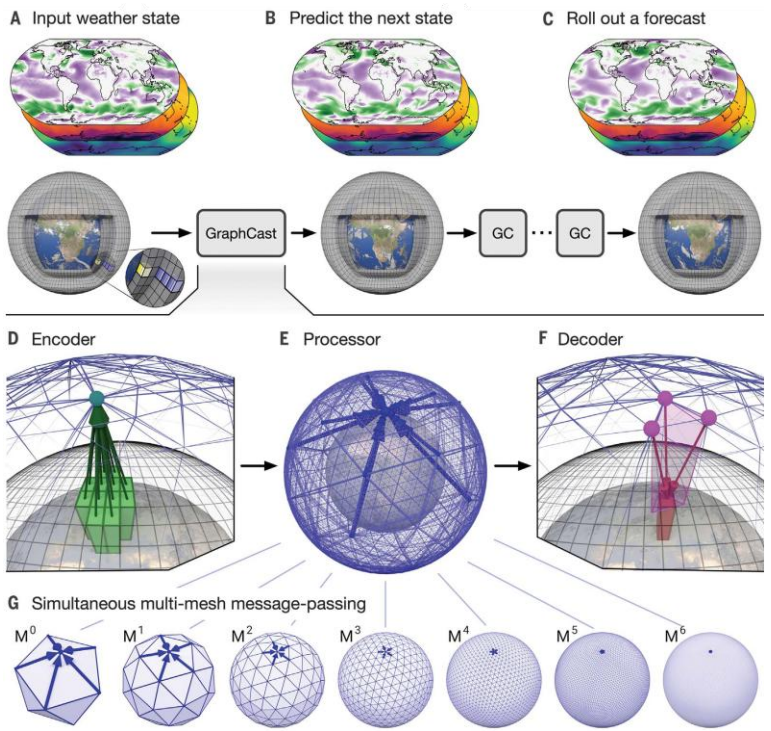
Improved protein structure prediction using potentials from deep learning



AlphaFold

Science

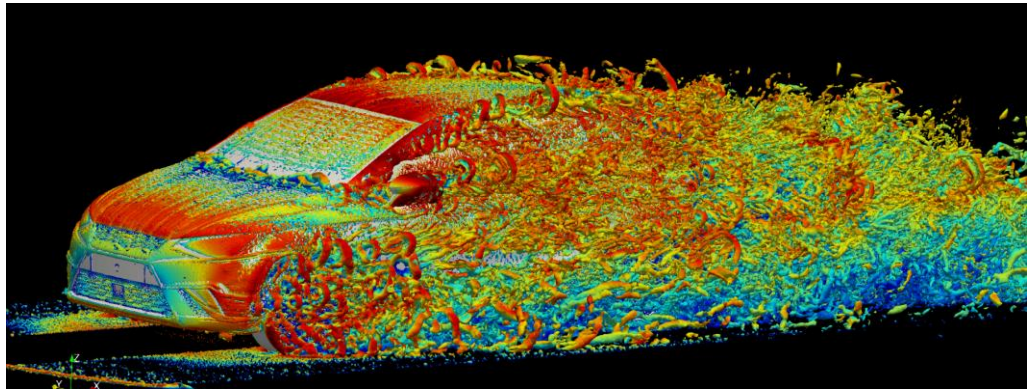
Learning skillful medium-range global weather forecasting



GraphCast

AI for Spatial and Spatio-temporal Data

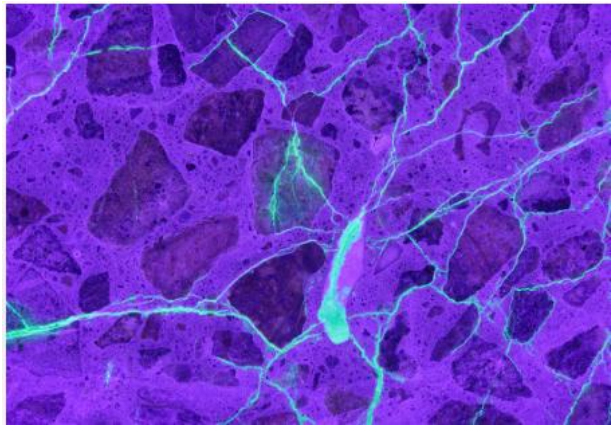
Complex Spatial-Temporal Data Is Common for Science & Engineering



Fluid dynamics

Climate/Weather Prediction

Satellite Remote Sensing



X-ray/electron microscopy

Industrial inspection

Synchrotron beamline

Multi-dimensional Images in Science & Engineering

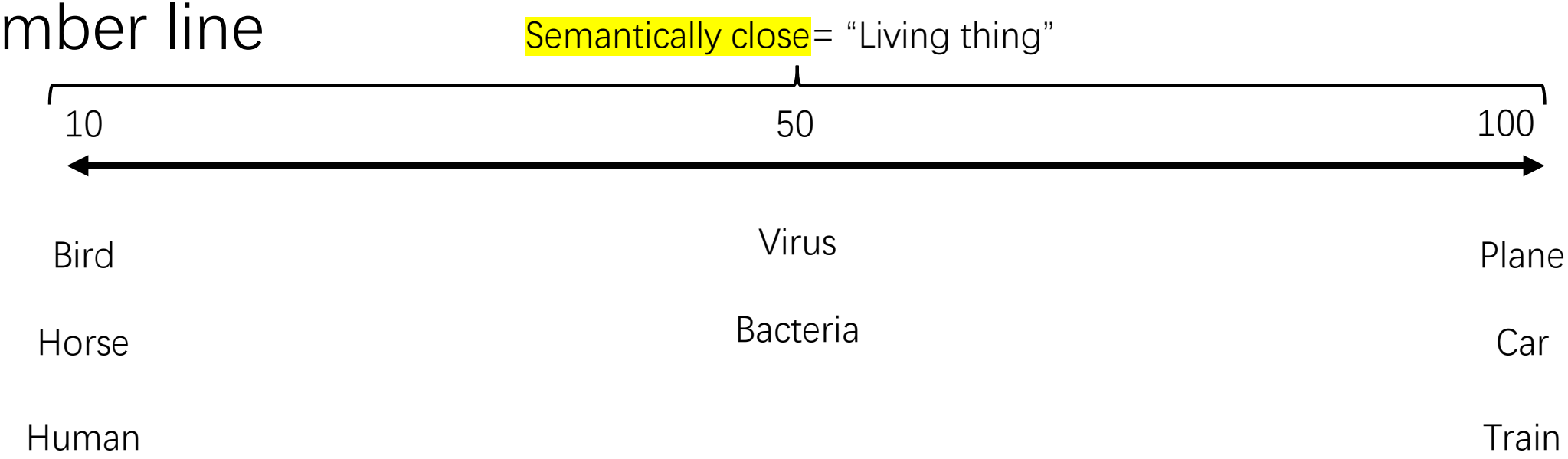
Type	Resolution	Tokens/Sample Patch = 162/163	Dataset Sizes	Dimensions (Channels)
Weather\Climate Simulations	$100s^3$ 10s channels (ERA5 dataset)	~ 300K	~10 PB	3 Spatial + 1 Temporal + (N Channels)
Satellite Images	$1000s^3$ 10s channels	~5M	~ 10s TB	2 Spatial + 1 Temporal + (N Channels)
Microscopic (Ex: Pathology)	$100K^2$	~100Ks (4x4 patch)	~ 10s TBs	2 Spatial + (1 or N Channels)
Video	$100s^2$ ~ Hours (24 f/s) (YouTube-8m)	~1M	~1 PB	2 Spatial + 1 Temporal + (N Channels)
X-Ray CT (Ex: SP-μCT)	~$8\text{-}12K^3$ >163 new beam	~1B	~100s TB	3 Spatial + (1 Channel)
MRI (Ex: dMRI)	~$4K^3$ (sub 5-micron)	~ 30M	~ 10s TB	3 Spatial + (N Channels)

What is a Transformer (LLMs)?

Self-attention TL;DR: Text and Embeddings

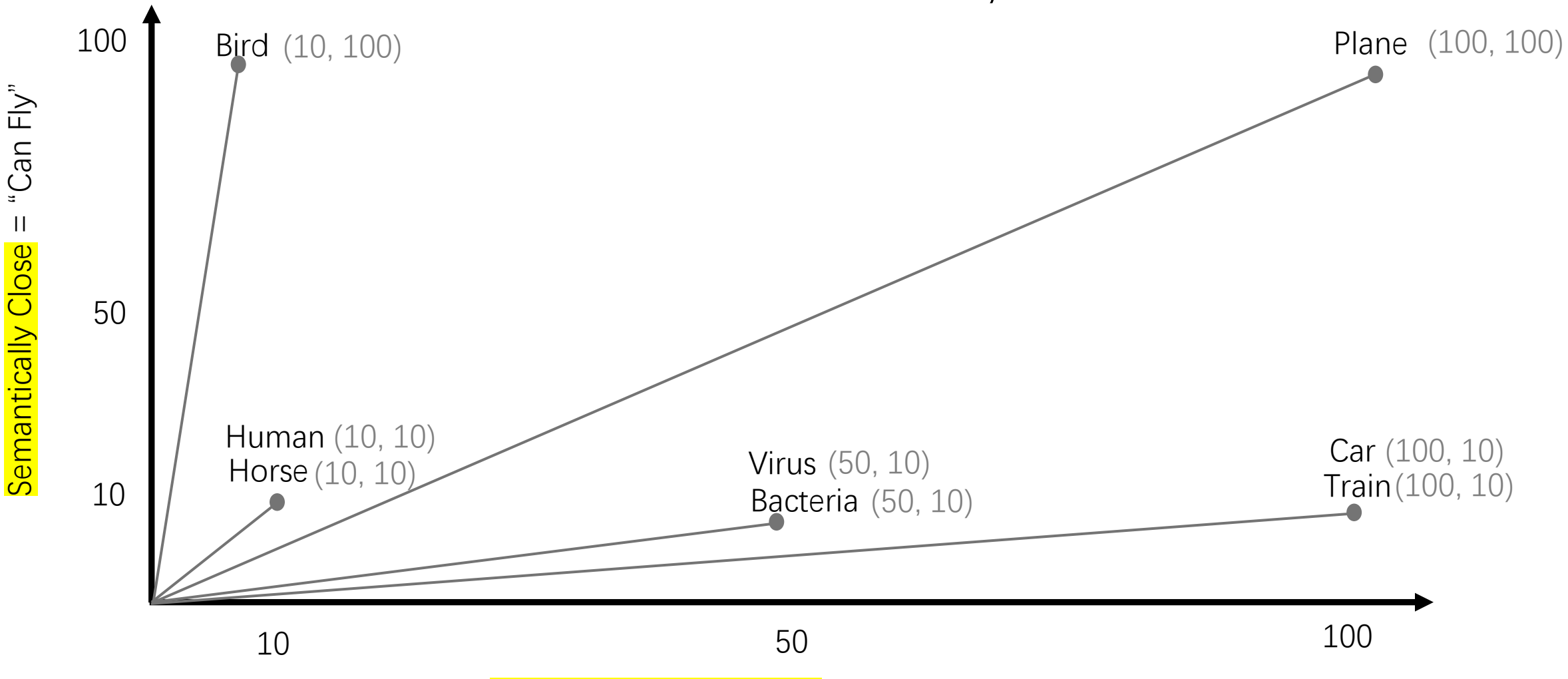
- Modeling language:
  - Find how words relate to and affect the meaning of other words

- Use a single number to represent each word, words semantically close in meaning are close to each other on the number line



Text and Embeddings

- But words can be close in one attribute, and not others



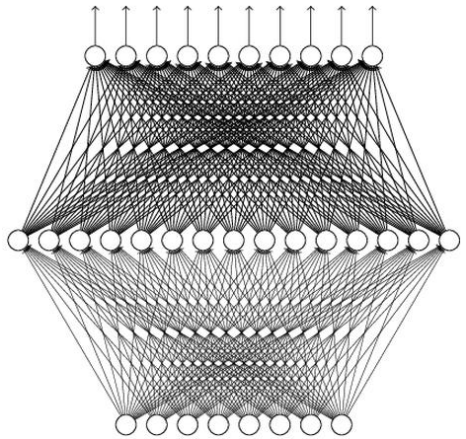
ViT and Inductive Bias

- Universal law of approximation

Any continuous function f

$$f : R^N \rightarrow R^M$$

Can be realized by a fully connected network with hidden layer(s)



Reference for the reason:
http://neuralnetworksanddeeplearning.com/chap4.html

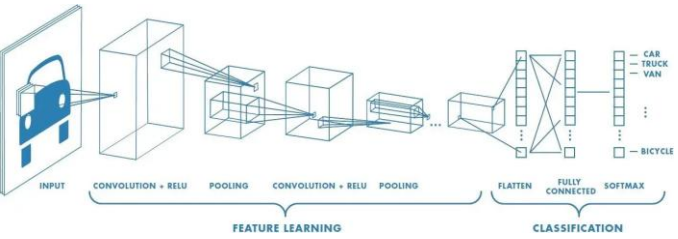
- Bound on approximation error ☑
  - Generalization error ☒
  - Optimizer error ☒

- FC Networks not enough (in Practice)

AI in Science: use ML pattern recognition capability to understand nature

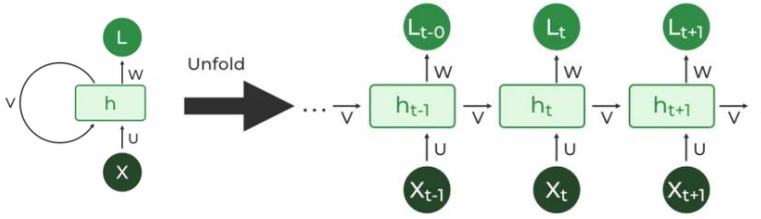
- Neural Networks → sophisticated pattern recognition
  - Add network elements to "work better" with the pattern in the data

Spatially-local Pattern



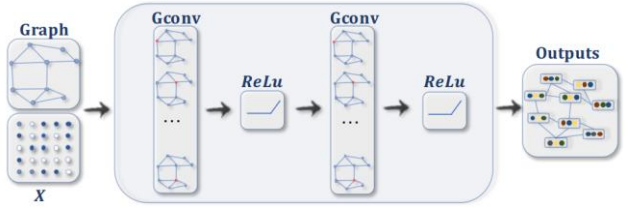
Convolution Neural Network

Temporal Pattern



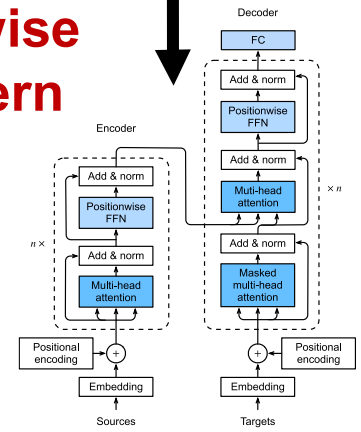
Recurrent Neural Network

Structural Relationships



Graph Neural Network

Pairwise Pattern



Transformers (w/ self-attention)

ViT and Inductive Bias

- From the Inductive bias POV
  - ViT might not seem to make much sense
- Receptive field is entire image
  - Every piece of the image attends to all other pieces of the image

- **Positives**
  - Model learns EVERYTHING → can ingest massive datasets

- **Negatives**
  - Model learns EVERYTHING → expensive; finds irrelevant patterns

ViT Issues vs. Text Transformer Issues

- Transformer consume sequence of tokens
  - Fitting to the nature of text

- Text tokens: atomic semantically distinct, rich in information,
  - Visual tokens: geometrically related and sparse in semantics

- **Loss of spatial hierarchy information becomes more pronounced**
  - When working with high-resolution or high-dimension images

ViT Challenges

1. Long sequence (when ingesting all image elements at high res)
2. Shifting bottleneck (elements outside the Transformer encoder)
3. Tokenization (managing temporal dimension)
4. Different parallelism in training (vs. text Transformer)
5. Multi-modality

Two Main Principles on Solving Real-world Problems

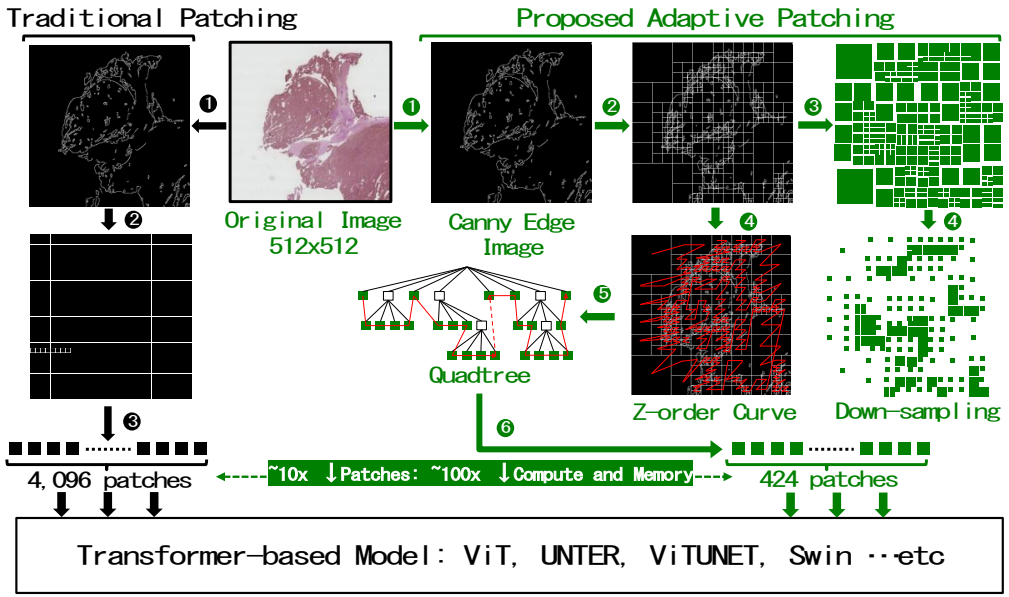
1. Start from Scientific Inquiry; Work Back to Solution

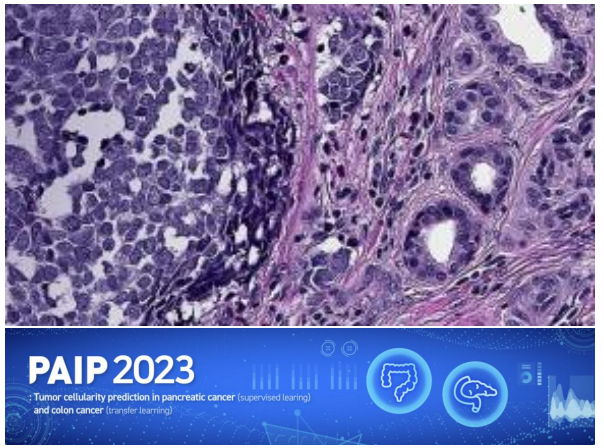
2. AI is a toolbox; always pick the right tool

Vision Transformers Use in Real-world Problems

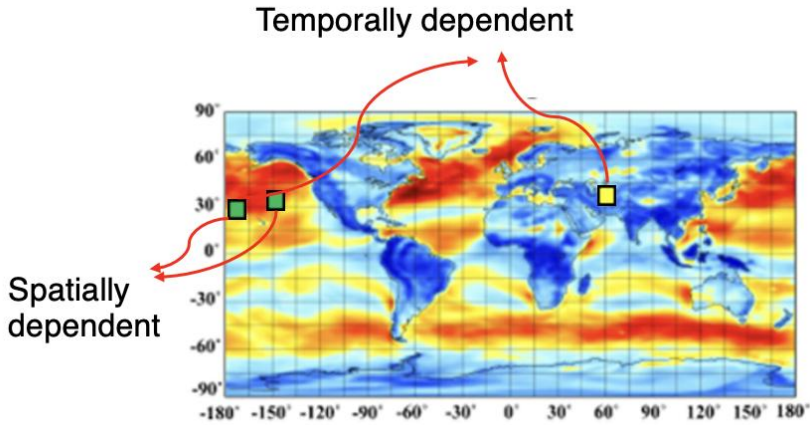
Vision Transformers used in production, Examples:

a) microscopic pathology, b) X-ray CT road samples, c) weather prediction

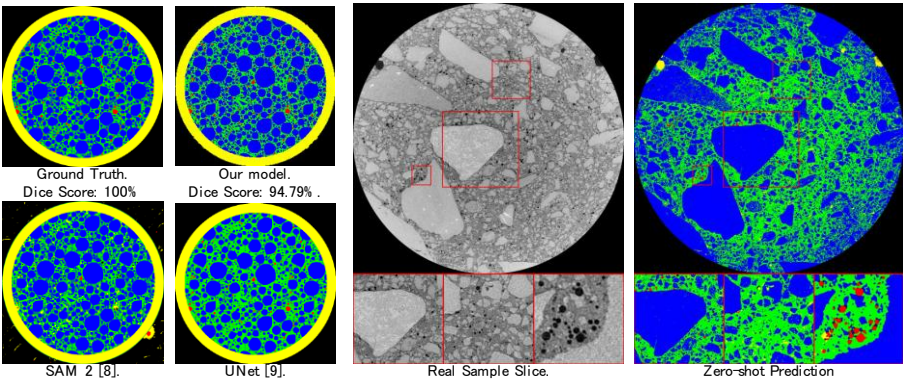


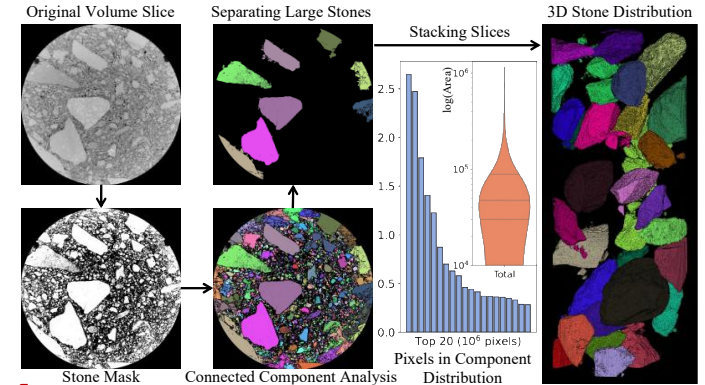


3

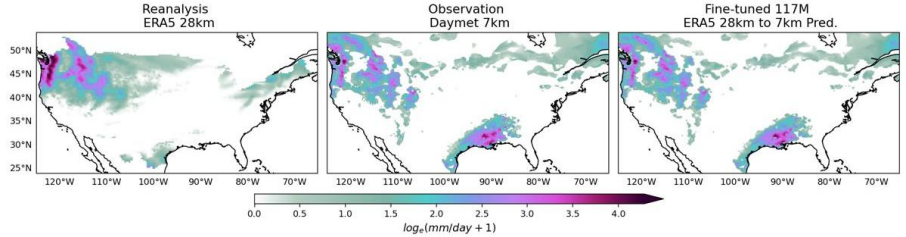


2





1



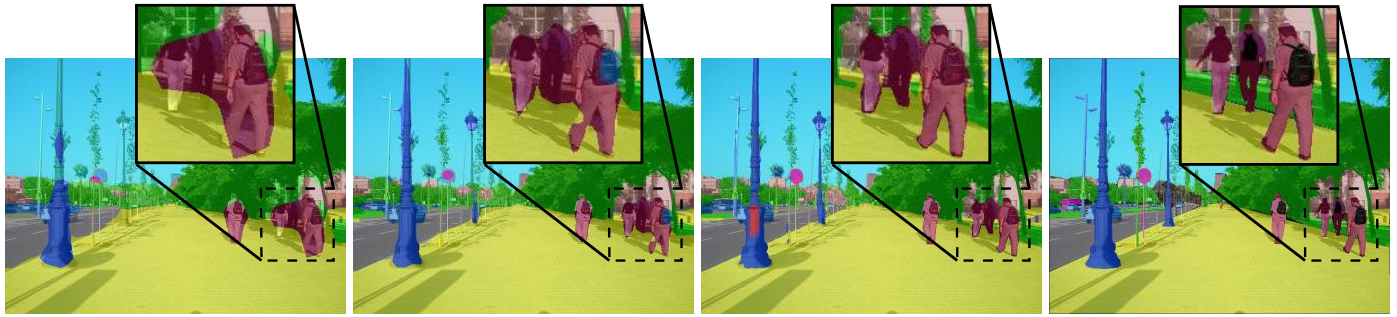
1 2 3

ViT Challenges

1. Long sequence (when ingesting all image elements at high res)
2. Shifting bottleneck (elements outside the Transformer encoder)
3. Tokenization (managing temporal dimension)
4. Different parallelism in training (vs. text Transformer)
5. Multi-modality

LLMs for 3D Segmentation

- **Using LLMs for Vision (ex: Vision Transformers)**
  - Because of self-attention, the receptive field is the entire image!
  - Split image to patches (ex: 16x16)
  - Feed patches to LLMs
- **Segmentation**
  - Larger patches → model learns global meaningful segmentation; produces poor boundaries
  - Smaller patches are qualitatively better
  - 4x4 patches for $4K^3$ 3D image = 1,000,000,000 tokens/image



(a) Patch size 32×32 (b) Patch size 16×16 (c) Patch size 8×8 (d) Ground Truth

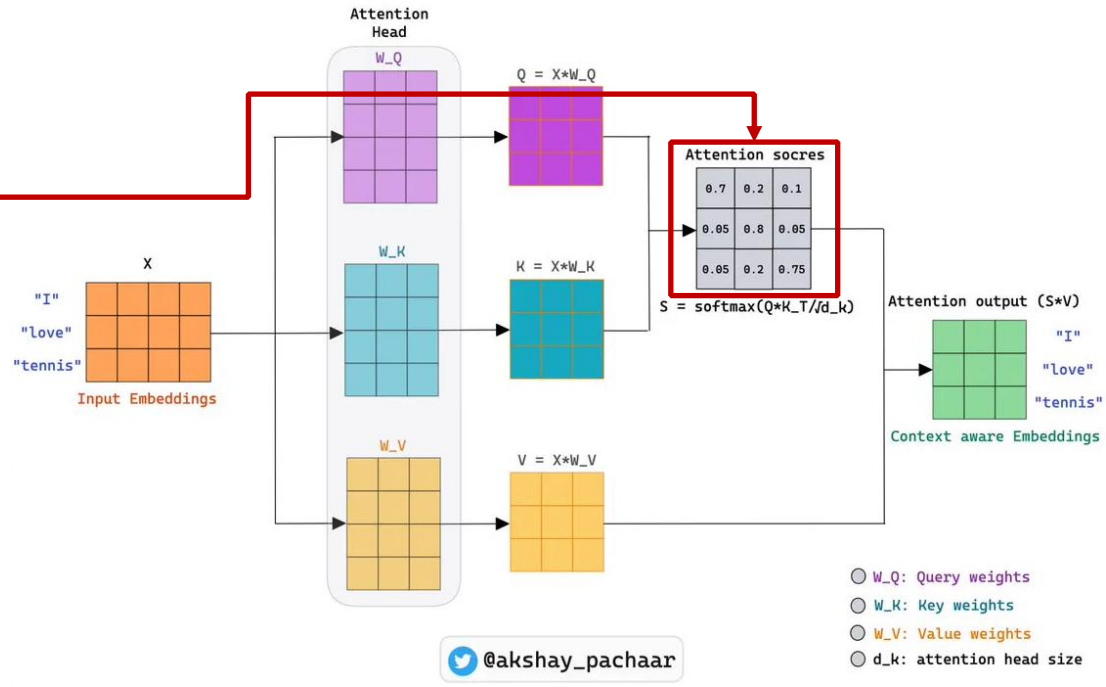
Impact of Model Patch Size on the Segmentation Maps*

* *Strudel et al. Segmenter: Transformer for Semantic Segmentation, ICCV'21 (https://arxiv.org/pdf/2105.05633.pdf)*

Longer Sequence: a Challenge

- The longer the sequence, the more the **context** that can be extracted
  - Ex: feeding an LLM entire books, library of papers, RAG, or **segmentation**
  - GPT-4-turbo → 128,000 tokens – GPT4-32k → 32,768 tokens (1 Token = ¾ Word)

- Compute and memory cost ∝ sequence^2

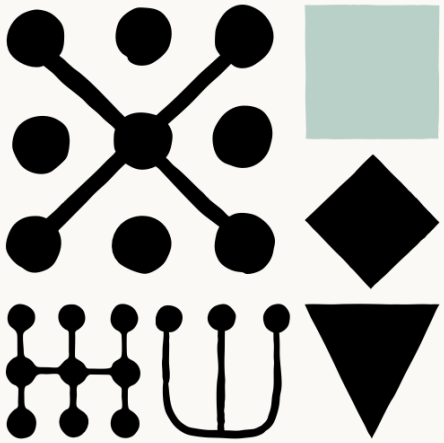


Methods for Longer Sequence

ANTHROP\C

Claude API Solutions Research Commitments Learn News Try Claude

Engineering at Anthropic



How we built our multi-agent research system

Published Jun 13, 2025

Our Research feature uses multiple Claude agents to explore complex topics more effectively. We share the engineering challenges and the lessons we learned from building this system.

"*Multi-agent architectures effectively scale token usage for tasks that exceed the limits of single agents.*"

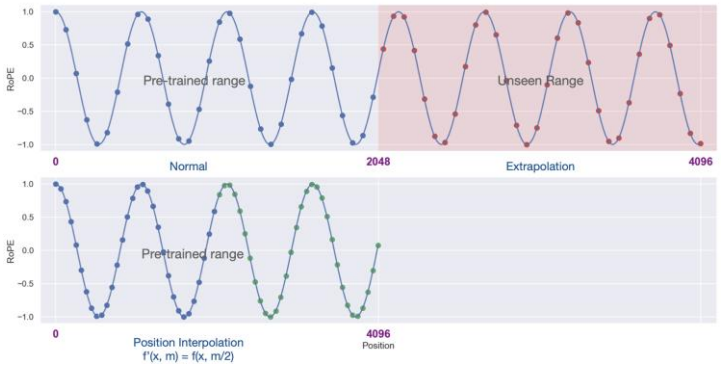
Compatibility of Longer Sequence Approaches in Training

	Alternative for Attention	Hierarchal Training	Attention Approximation	Cache Blocking	Reduce Tokens	Sequence Parallelism
Alternative for Attention		?	✗	?	✓	?
Hierarchal Training	?		✓	✓	✓	✓
Attention Approximation	✗	✓		?	✓	✗
Cache Blocking	?	✓	?		✓	✓
Reduce Tokens	✓	✓	✓	✓		✓
Sequence Parallelism	?	✓	✗	✓	✓	

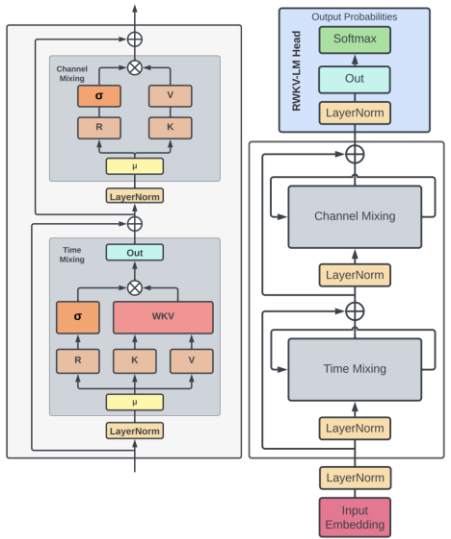
Methods for Longer Sequence

➢Alternative Mechanism for Attention

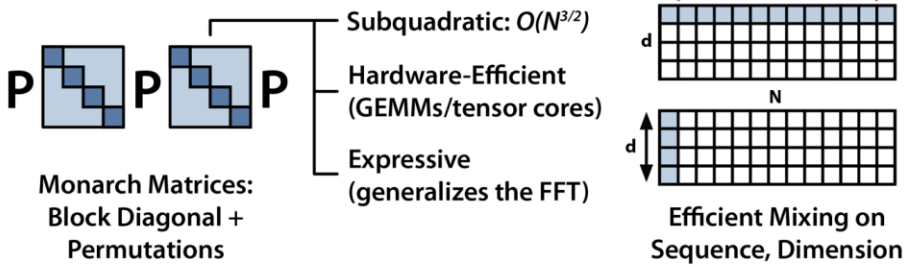
Positional Interpolation: downscale vs. extrapolating (Extend window size)



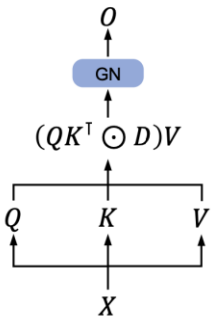
RWKV: Transformer in training; RNN in inference (Linear attention)

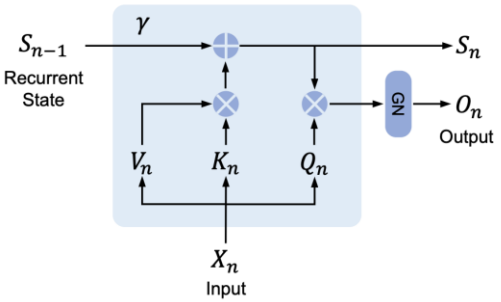


Monarch: use convolution to compute Attention (FFT for convolution)



RetNet: a retention mechanism for attention modeling (parallel recurrency)

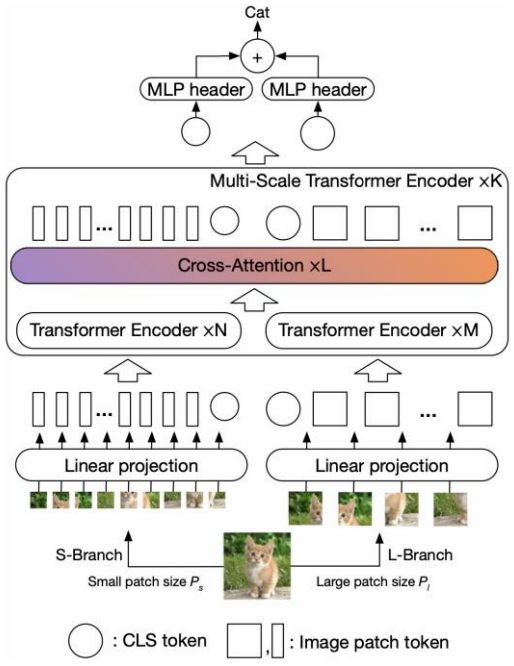




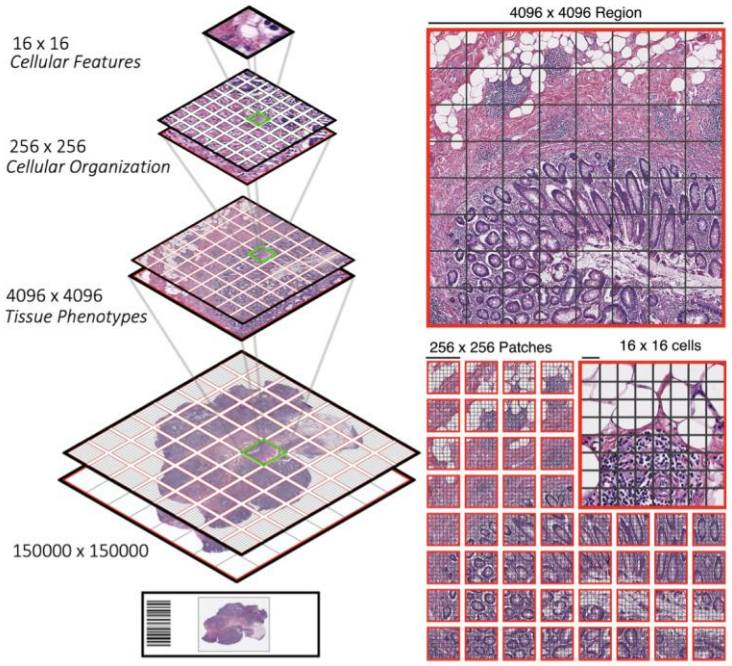
Methods for Longer Sequence

- Hierarchal Training
  - Train multiple transformers at different levels of abstraction
  - The transformer at the lowest abstraction level trains on the shortest sequence segments.
  - The transformer at the next higher level uses the previous level outputs as additional input to train on longer segments.

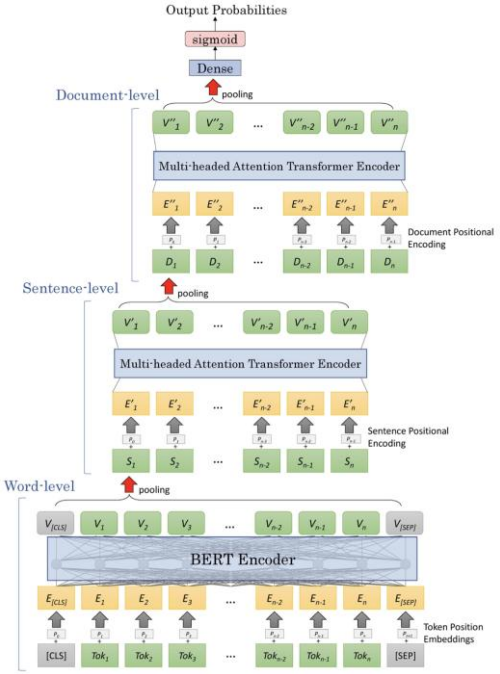
CrossViT



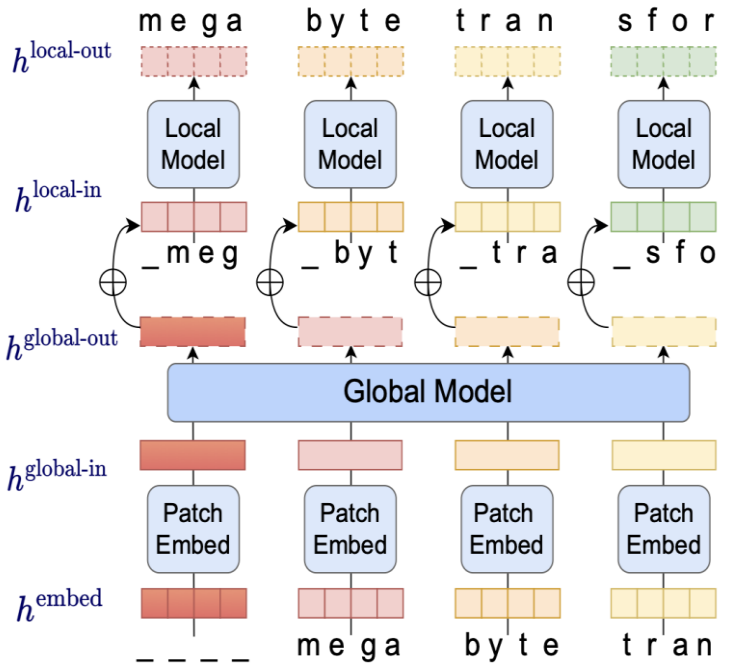
Hierarchal ViT



Three-level Transformer



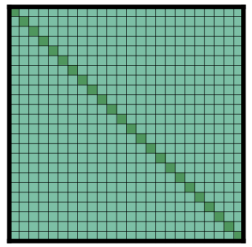
MEGABYTE

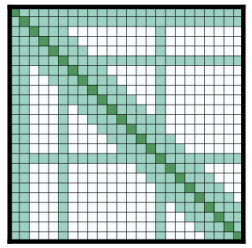


Methods for Longer Sequence

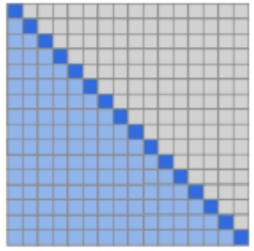
- Attention Approximation
  - Approximate self-attention operation through sparse sampling, lox-rank approx., infrequent update …
  - Note: “sparse” when talking about LLMs now mean **sparse sequence**, not sparse model

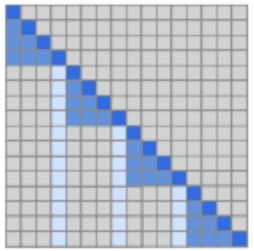
Longformer



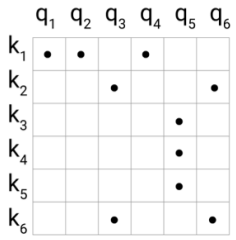


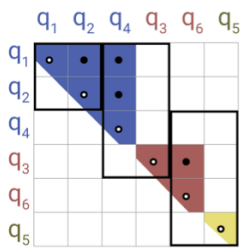
Sparse Transformer



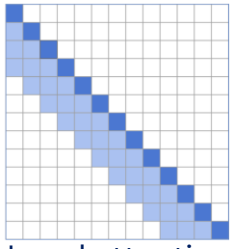


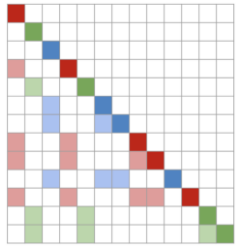
Reformer (Hash attention)



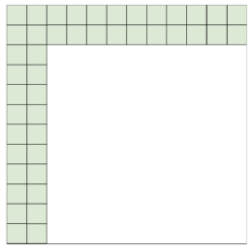


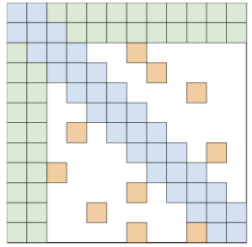
Routing Transformer



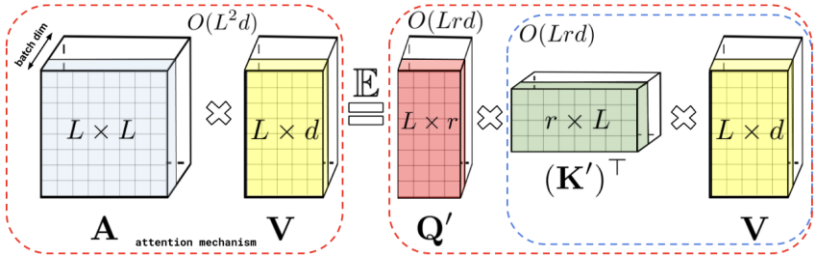


Big Bird





Performers



Linear Transformers

$$A_l(x) = V' = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V. \quad \Rightarrow \quad V_i' = \frac{\sum_{j=1}^{N} \text{sim}(Q_i, K_j) V_j}{\sum_{j=1}^{N} \text{sim}(Q_i, K_j)}.$$

In-frequent Update

$$s_i = \text{dot}(q, k_i), \qquad s_i' = e^{s_i}, \qquad \text{attention}(q, k, v) = \frac{\sum_i v_i s_i'}{\sum_j s_j'}.$$

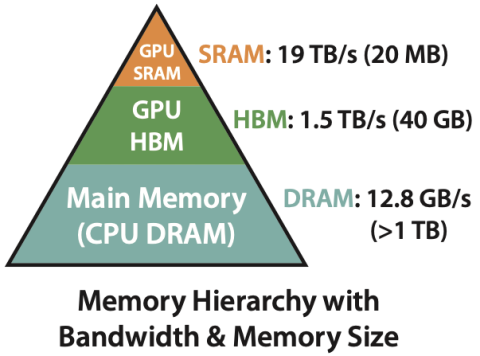
LazyFormer

Remove dropout in self-attention.
Wider Layers.

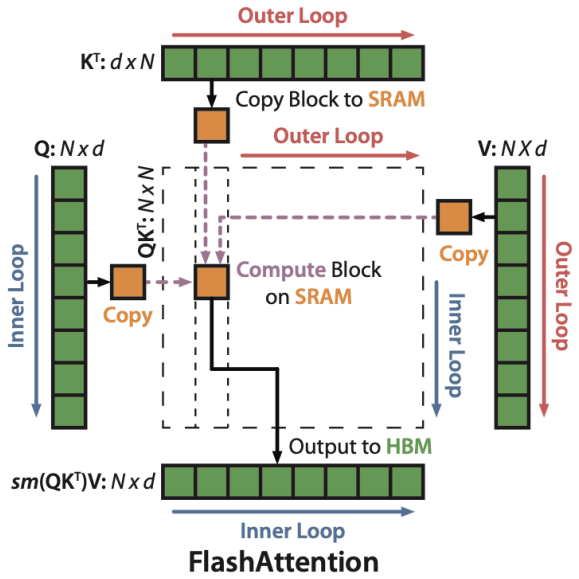
Methods for Longer Sequence

- Cache Blocking
  - No approximation
  - Can support longer sequences by blocking the attention matrix in scratchpad memory
  - Aggregate amount of work stays the same → **support longer sequence, but not faster**

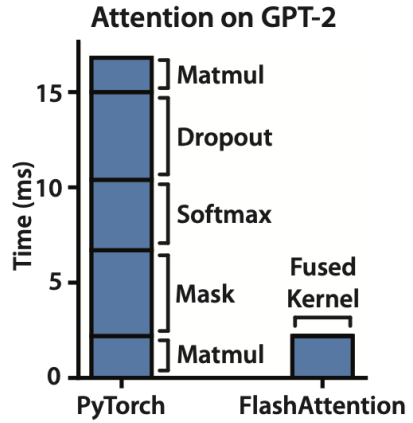
FlashAttention



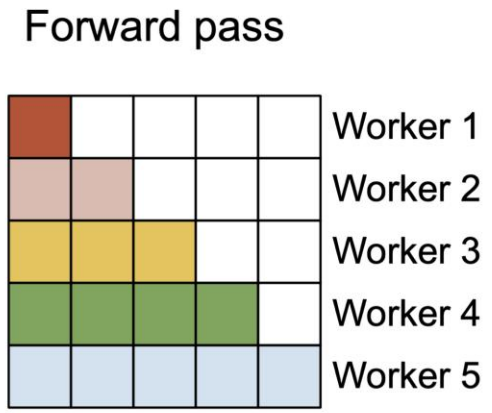
Memory Hierarchy with Bandwidth & Memory Size

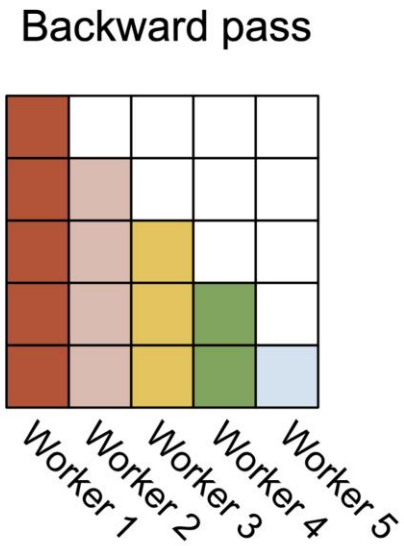


FlashAttention



FlashAttention2

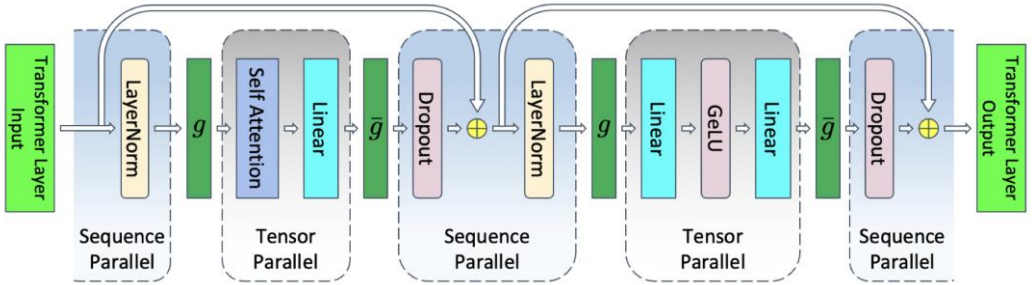




Methods for Longer Sequence

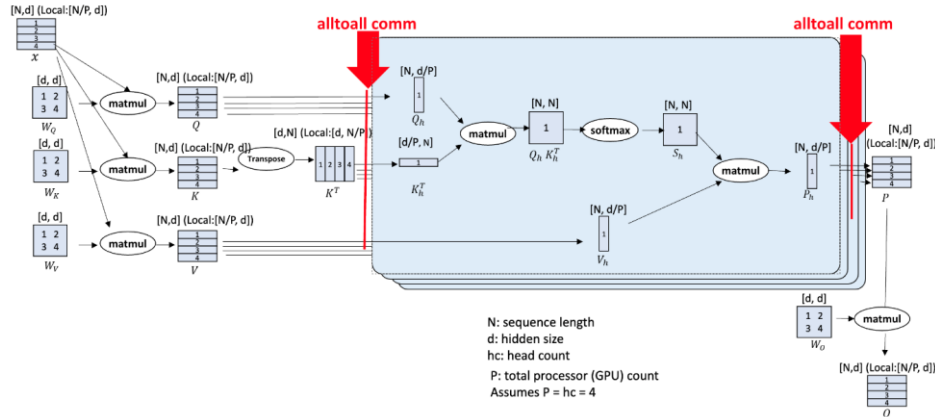
- Sequence/context Parallelism
  - Distributing long sequences among GPUs as short contiguous segments
  - Communication overhead due to token inter-dependence

Nvidia Sequence Parallelism



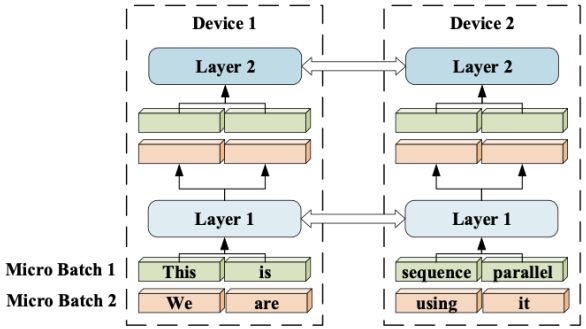
*Not really sequence parallelism; parallelizes Dropout and LayerNorm

MS DeepSpeed-Ulysses



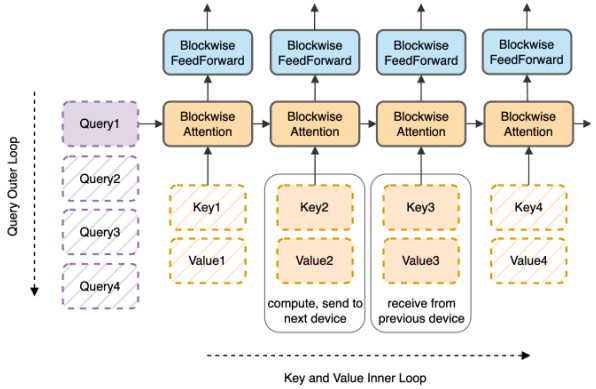
*Not really sequence parallelism; full attention is still at each worker

Long Sequence

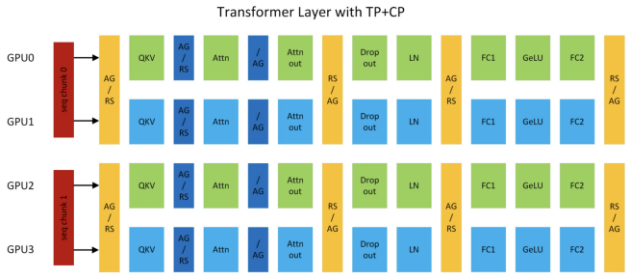


*No significant speedup

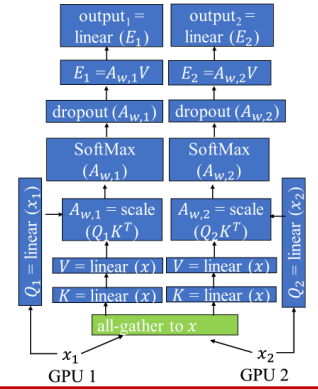
Ring Attention



Nvidia Context Parallel



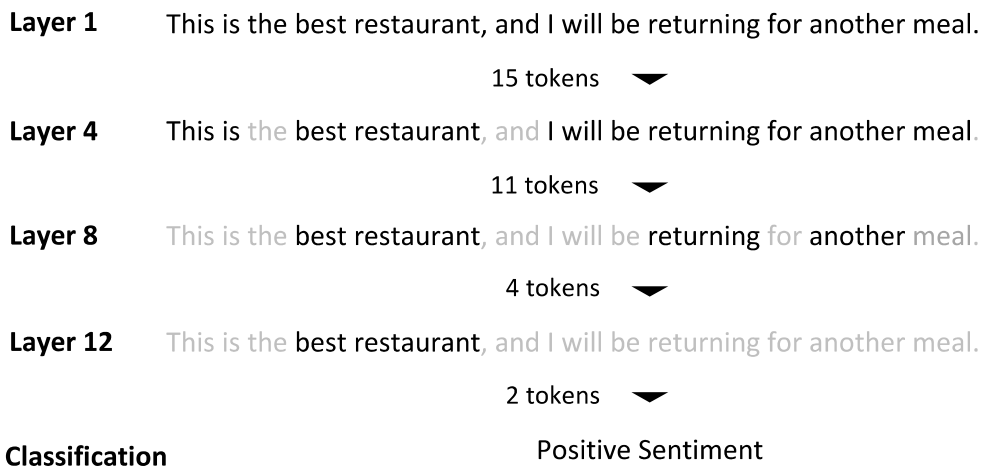
Fully Distributed Sequence



Methods for Longer Sequence

- Reduce Amount of Tokens
  - Current practice: divide input to tokens, feed all tokens to the model
  - Feed the model less tokens: BEFORE tokens are ingested or DURING passing through the model

(Learned) Token Pruning

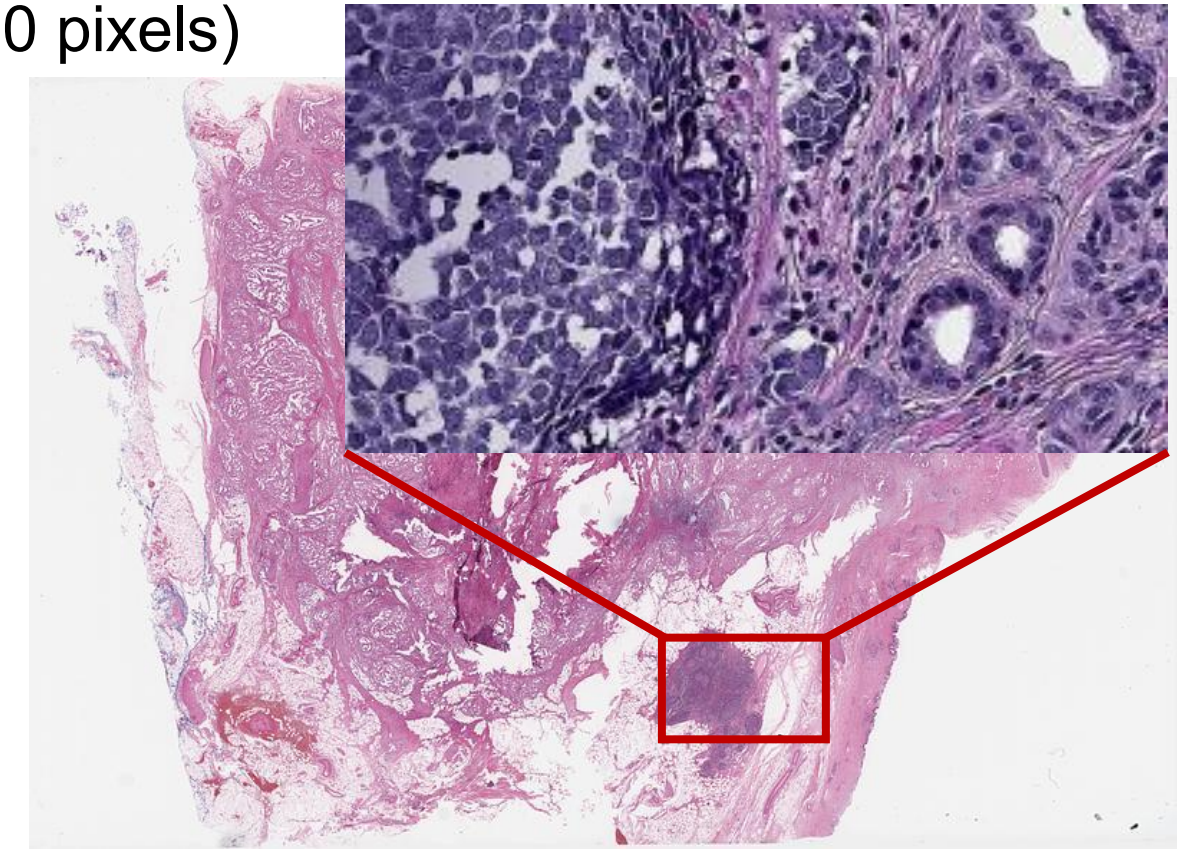


Adaptive Patching (ViT)

Patch the images in a ”smarter” way

Tumor Cellularity Prediction in Pancreatic Cancer and Colon Cancer

- Very high resolution (up to 100,000 x 100,000 pixels)
  - Used in pathology
  - Ex: PAIP dataset
    - Pancreas
    - **Diagnostic:** Perineural Invasion



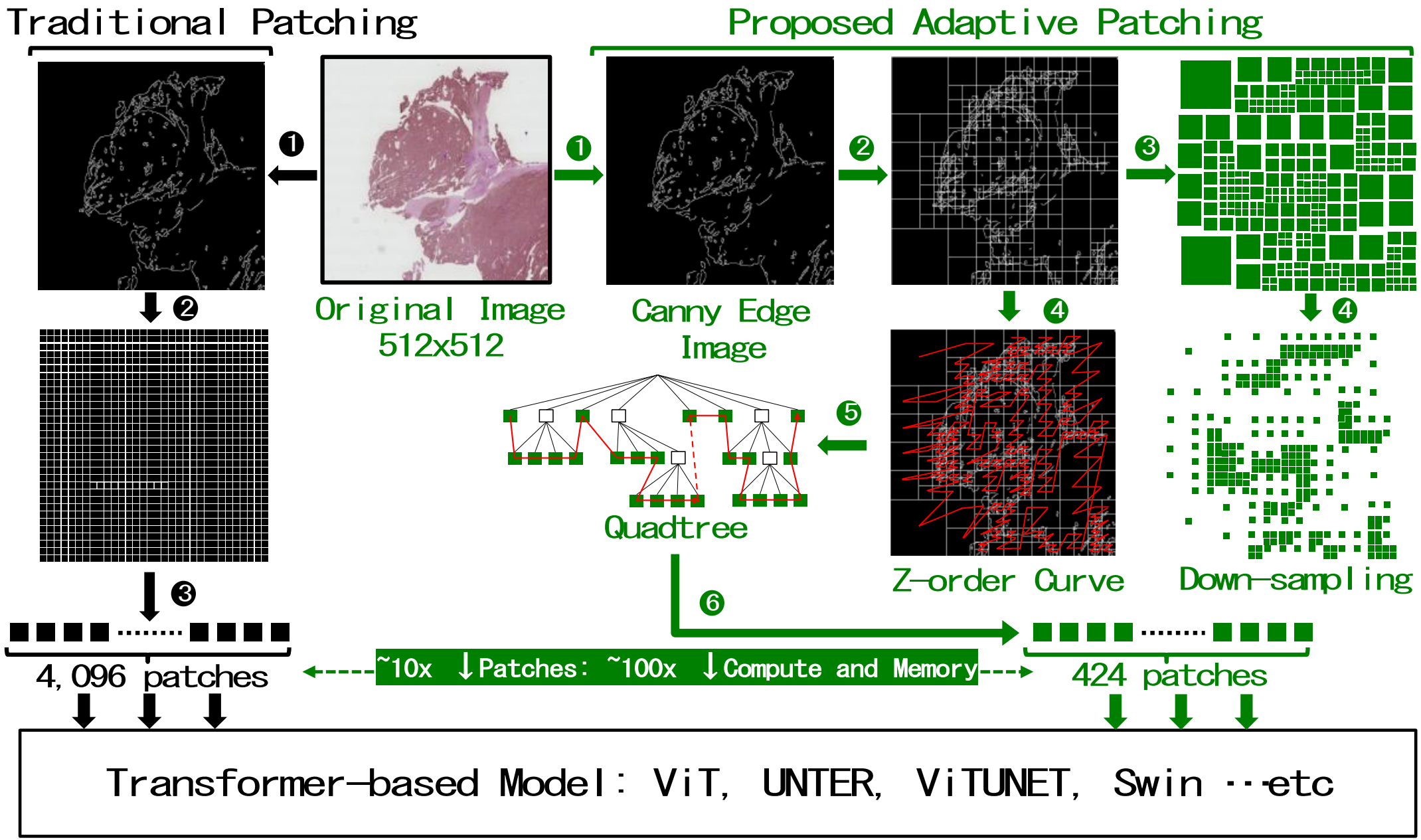
- Segmentation with Vision Transformer (ViT)

- Challenge:

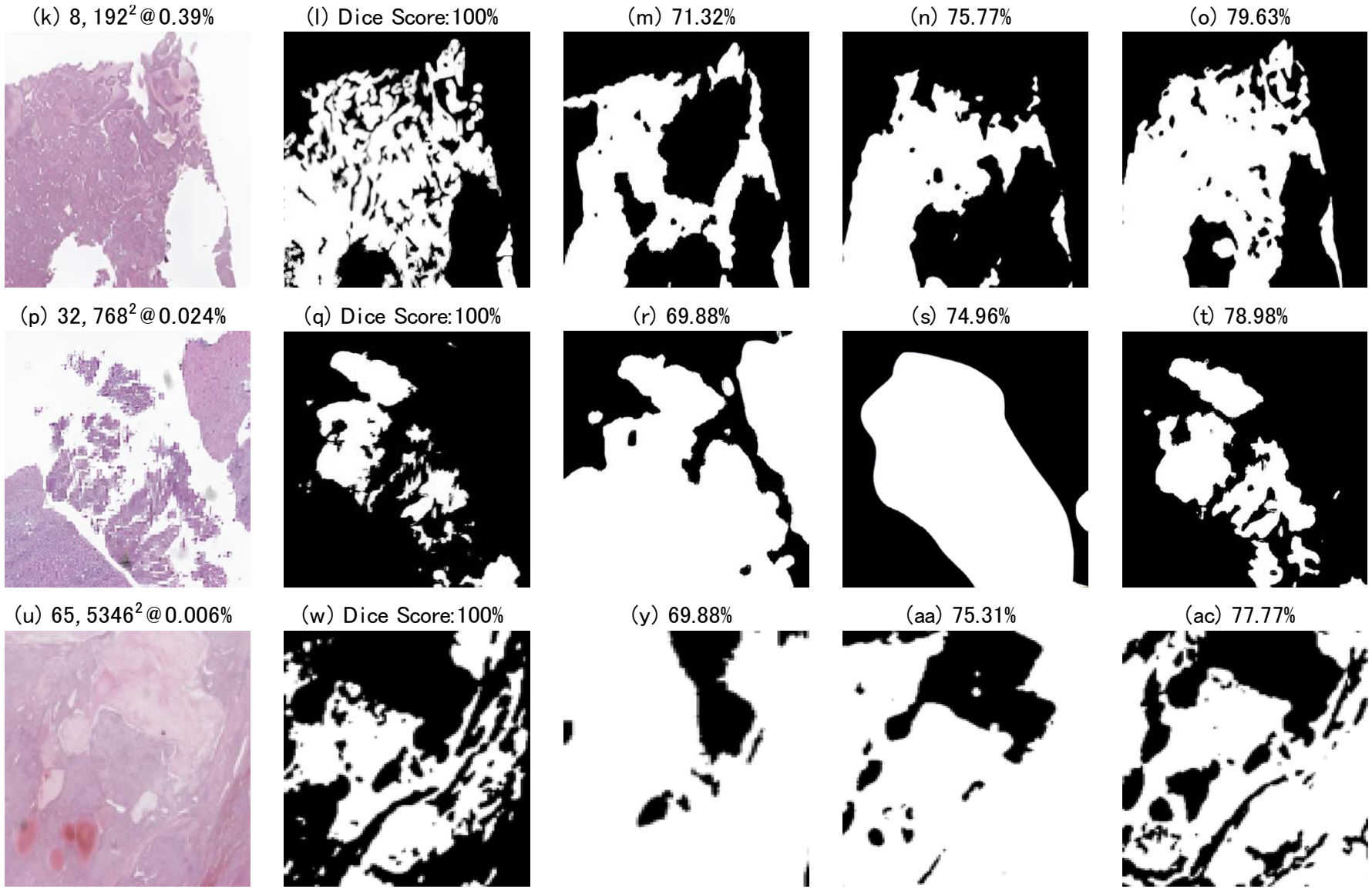
PAIP 2023: Tumor cellularity prediction in pancreatic cancer and colon cancer (transfer learning)

* E. Zhang et al. Adaptive Patching for High-resolution Image Segmentation with Transformers, SC'24

Adaptive Patching: Ingest Only the Data that Matters



Tumor Cellularity Prediction in Pancreatic Cancer and Colon Cancer



Results: Speedup

Resolution	Model-Patch	Sec/Image	Sequence Length	Quadtree Depth	Dice Score (%)	Speedup (Sec/Image)	Speedup (Time to Convergence)
512×512 1 GPU	**APF-4** UNETR-4	0.06495 0.4863	1,024 16,384	7 -	77.88 77.31	7.48×	12.71×
$1,024 \times 1,024$ 8 GPUs	**APF-8** UNETR-8	0.14284 1.0863	1,024 16,384	7 -	75.63 75.72	7.6×	12.92×
$4,096 \times 4,096$ 128 GPUs	**APF-16** UNETR-32	0.32231 1.8613	2,116 16,384	8 -	75.74 75.77	5.77×	9.8×
$8,192 \times 8,192$ 256 GPUs	**APF-16** UNETR-64	1.1613 2.6618	2,116 16,384	9 -	76.13 75.27	2.29×	3.89×
$16,384 \times 16,384$ 512 GPUs	**APF-32** UNETR-128	1.7613 5.1179	1,024 16,384	9 -	75.92 75.89	2.9×	4.93×
$32,768 \times 32,768$ 1024 GPUs	**APF-32** UNETR-256	2.1567 8.1896	2,116 16,384	10 -	75.32 74.96	3.79×	6.44×
$65,536 \times 65,536$ 2048 GPUs	**APF-32** UNETR-512	5.733 13.218	4,096 16,384	11 -	75.82 75.31	2.3×	3.91×

Results: Quality

- Since we can reduce the sequence length
  - We could increase the patch size, get better results (for the same compute budget)

Resolution	Model	Patch	GPUs	Sec/Image/GPU	Depth	Sequence Length	Dice Score	Dice Improvement
$8,192 \times 8,192$	**APF (+UNTER)**	2	256	2.3314	12	10,609	**79.56**	5.70%
		4	256	2.1314	11	8,464	78.31	
		8	128	1.7867	10	4,096	77.61	
		16	64	1.1613	9	2,116	76.13	
	UNETR	64	256	2.6618	-	16,384	75.27	
	TransUNet	-	256	2.3678	-	-	70.89	
	U-Net	-	32	1.2858	-	-	63.21	
$16,384 \times 16,384$	**APF (+UNTER)**	2	512	4.8792	13	16,384	**80.62**	6.23%
		4	256	3.1231	12	8,464	79.31	
		8	256	1.8574	11	4,096	78.84	
		16	128	1.6421	10	2,116	77.43	
	UNETR	128	512	5.1179	-	16,384	75.89	
	TransUNet	-	512	6.1296	-	-	70.46	
	U-Net	-	256	2.7825	-	-	62.97	
$32,768 \times 32,768$	**APF (+UNTER)**	4	1024	7.8916	13	16,384	**78.98**	5.36%
		8	512	6.1792	12	8,464	78.31	
		16	512	4.1685	11	4,096	77.61	
		32	256	2.1567	10	2,116	76.13	
	UNETR	256	1024	8.1896	-	16,384	74.96	
	TransUNet	-	1024	10.001	-	-	69.88	
	U-Net	-	512	4.2714	-	-	61.38	
$65,536 \times 65,536$	**APF (+UNTER)**	8	2048	12.697	13	16,384	**77.77**	3.27%
		16	1024	8.793	12	8,464	76.11	
		32	512	5.733	11	4,096	75.41	
		64	256	3.961	10	2,116	75.13	
	UNETR	512	2048	13.218	-	16,384	75.31	
	TransUNet	-	2048	14.3516	-	-	67.67	
	U-Net	-	1024	5.961	-	-	59.69	

AI for Cone-beam X-ray Computed Tomography

- X-ray CT widely used
- Current Generation X-ray CT rely on cone-beam scanners
- Higher quality; high-resolution real-time distributed reconstruction is intractable
- Use ViT to do geometry correction to make the real-time reconstruction tractable
  - $4K^3$ in 16 seconds and $8K^3$ in a few minutes (on 1,024 GPUs)

* P. Chen, M. Wahib et al. Scalable FBP decomposition for cone-beam CT reconstruction, SC'21

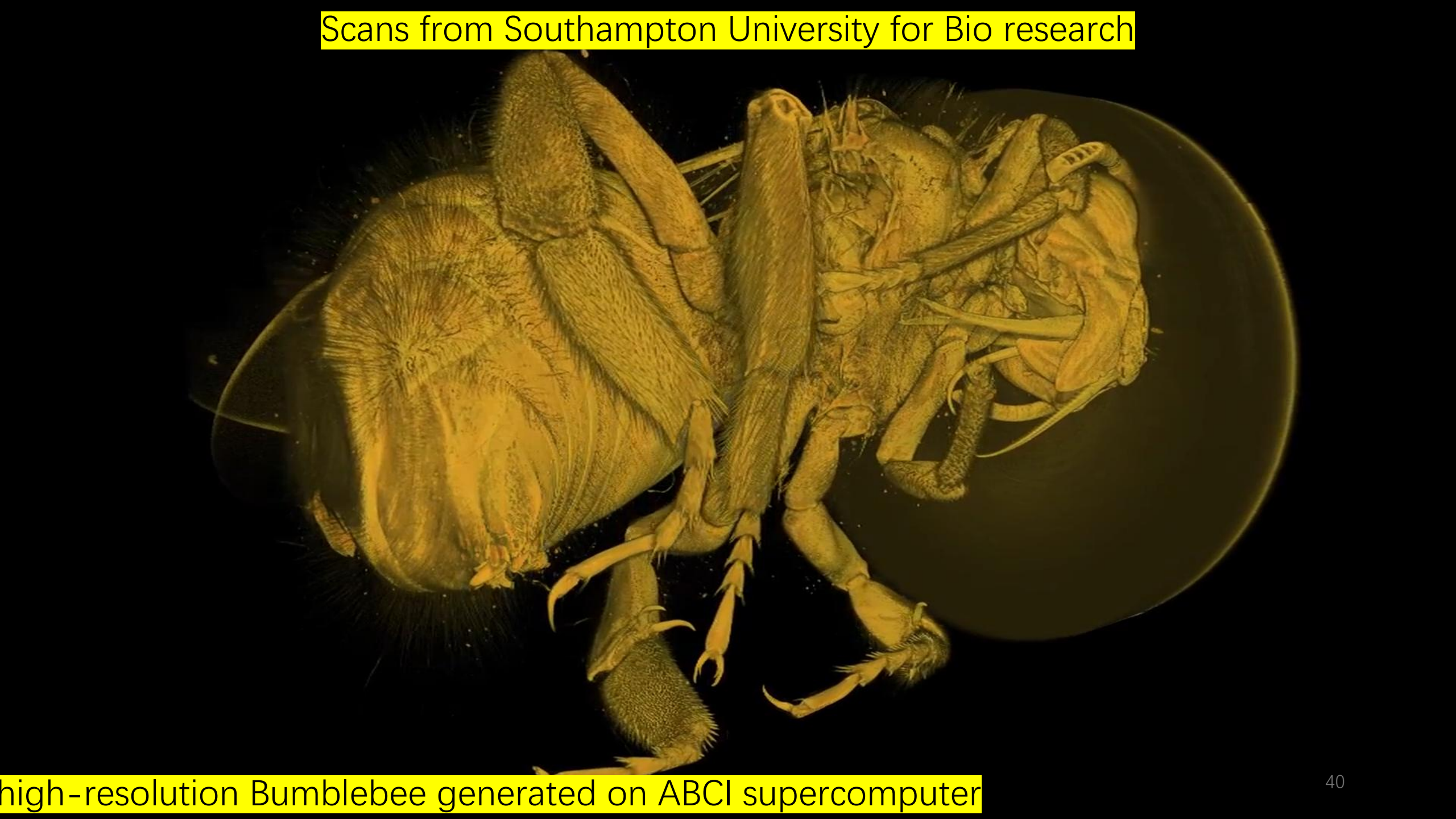
Scans from Southampton University for Bio research

high-resolution Bumblebee generated on ABCI supercomputer

Scans from Southampton University for Bio research

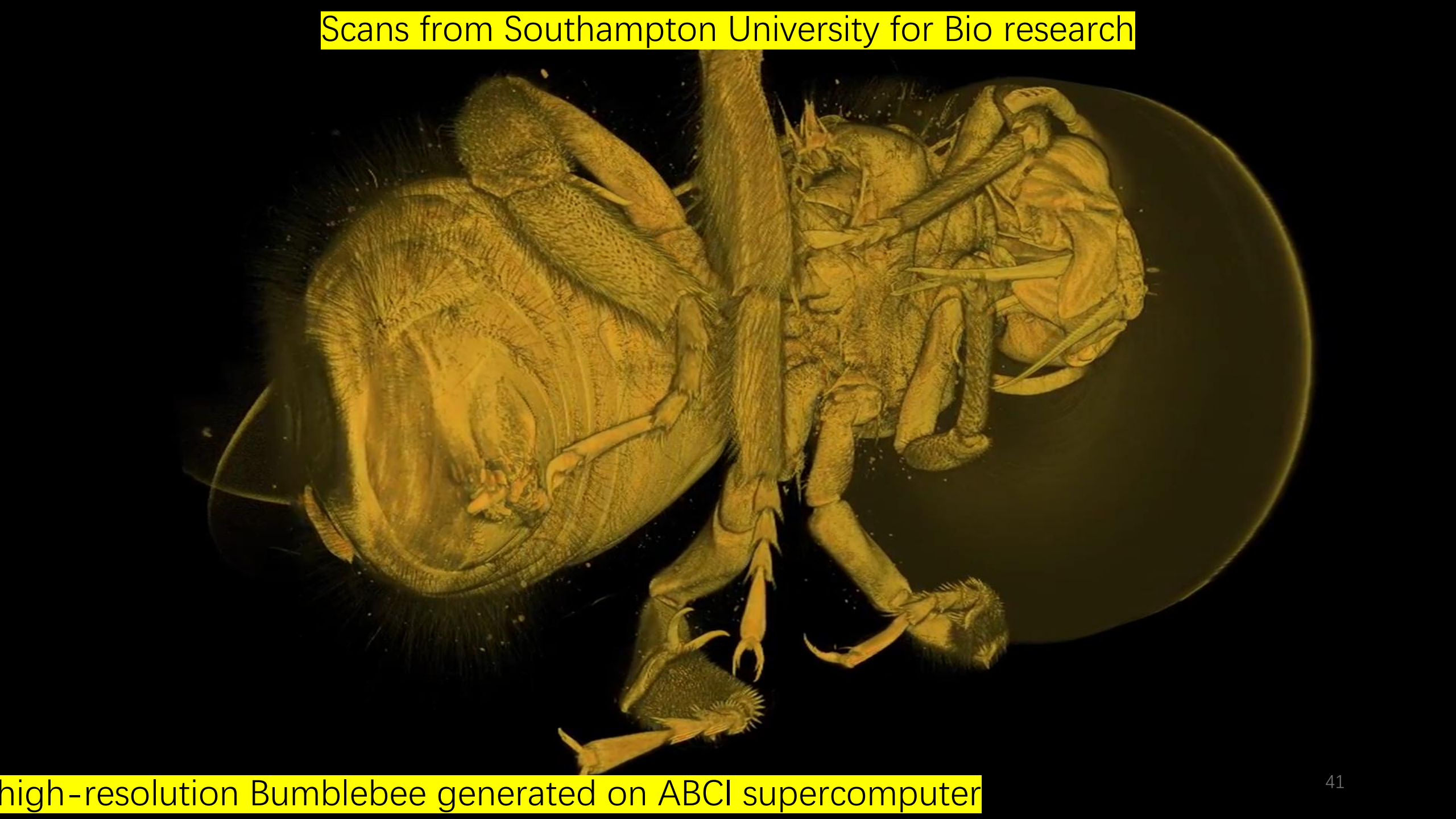
high-resolution Bumblebee generated on ABCI supercomputer

Scans from Southampton University for Bio research



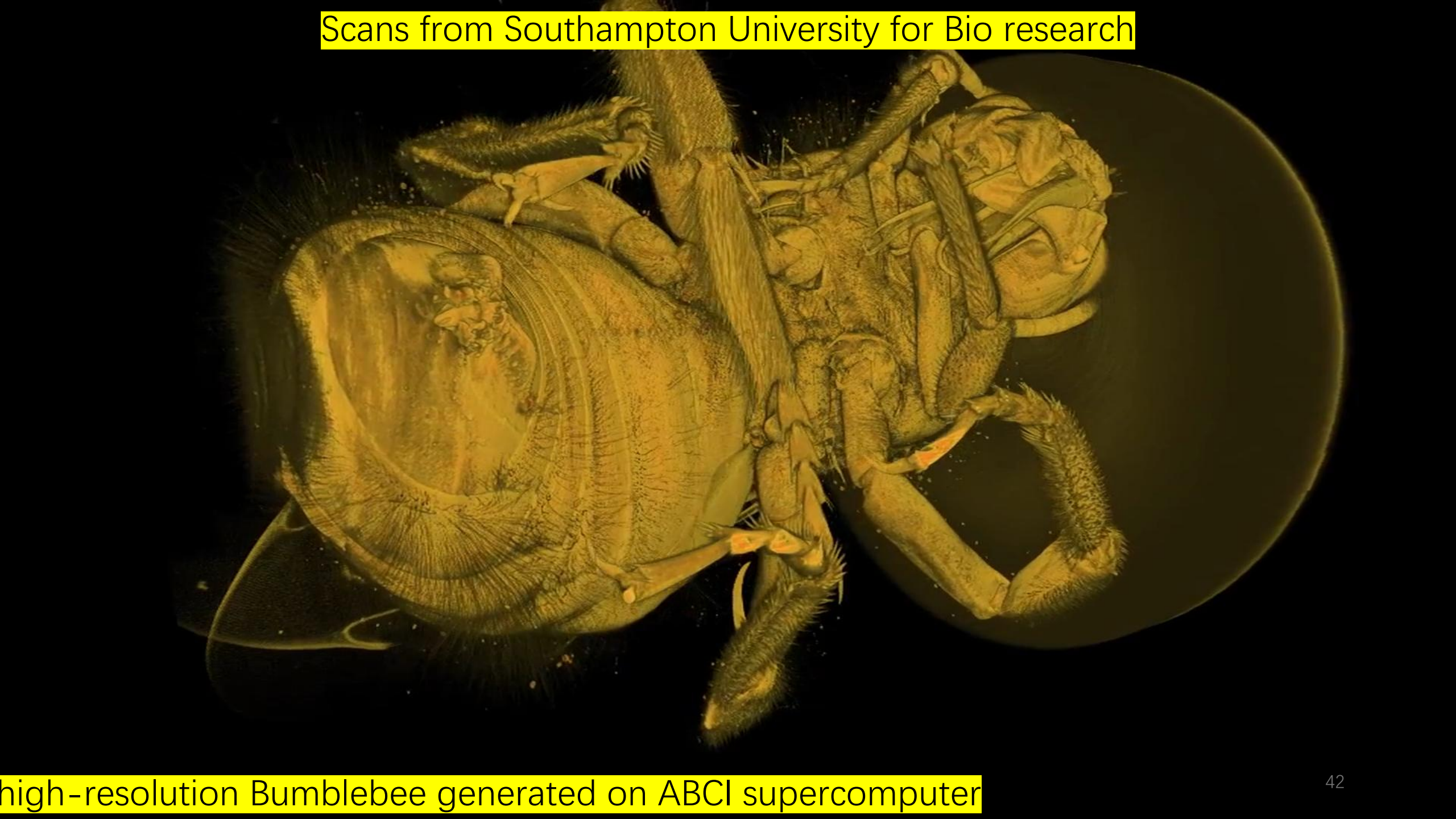
high-resolution Bumblebee generated on ABCI supercomputer

Scans from Southampton University for Bio research

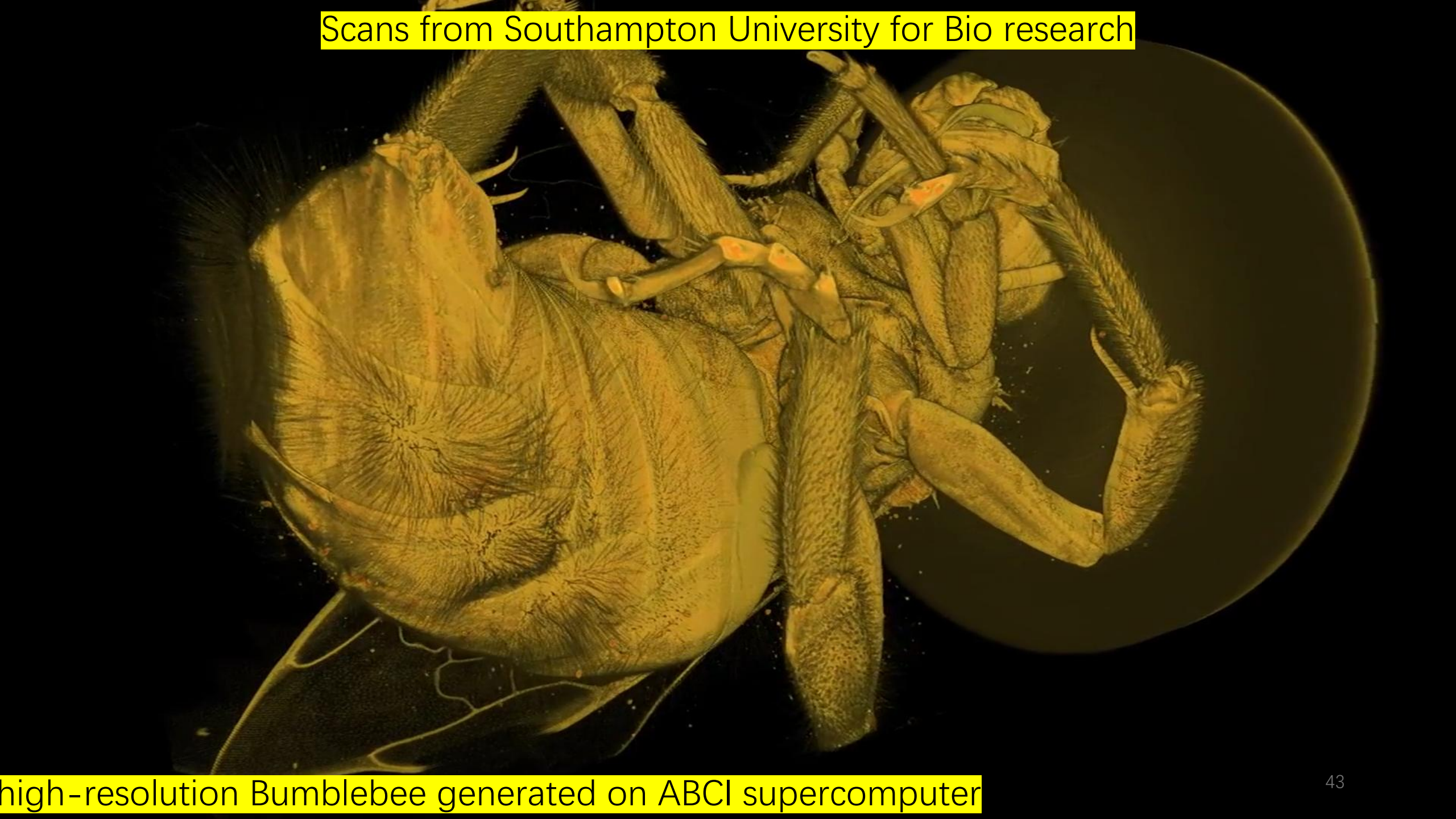


high-resolution Bumblebee generated on ABCI supercomputer

Scans from Southampton University for Bio research



high-resolution Bumblebee generated on ABCI supercomputer


Scans from Southampton University for Bio research
high-resolution Bumblebee generated on ABCI supercomputer


Scans from Southampton University for Bio research
high-resolution Bumblebee generated on ABCI supercomputer


Scans from Southampton University for Bio research
high-resolution Bumblebee generated on ABCI supercomputer


Scans from Southampton University for Bio research
high-resolution Bumblebee generated on ABCI supercomputer


Scans from Southampton University for Bio research
high-resolution Bumblebee generated on ABCI supercomputer


Scans from Southampton University for Bio research
high-resolution Bumblebee generated on ABCI supercomputer


Scans from Southampton University for Bio research
high-resolution Bumblebee generated on ABCI supercomputer

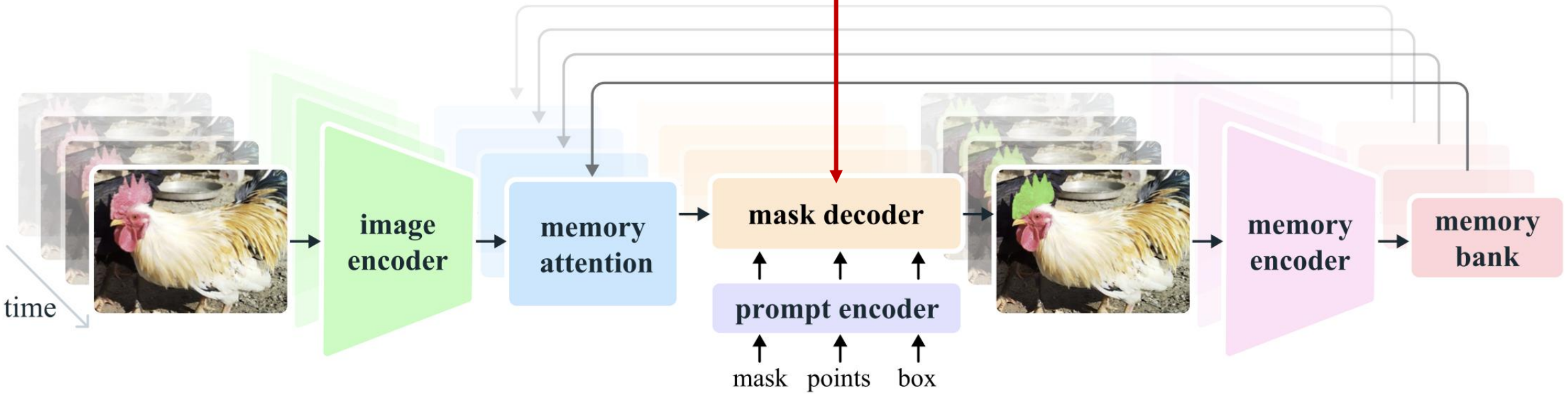

Scans from Southampton University for Bio research
high-resolution Bumblebee generated on ABCI supercomputer

ViT Challenges

1. Long sequence (when ingesting all image elements at high res)
2. Shifting bottleneck (elements outside the Transformer encoder)
3. Tokenization (managing temporal dimension)
4. Different parallelism in training (vs. text Transformer)
5. Multi-modality

SoTA Models Overwhelmed by High-resolution

- In Segmentation
  - We must map encoded features back to full-resolution pixel predictions
  - CNN mask decoder (too much memory required)



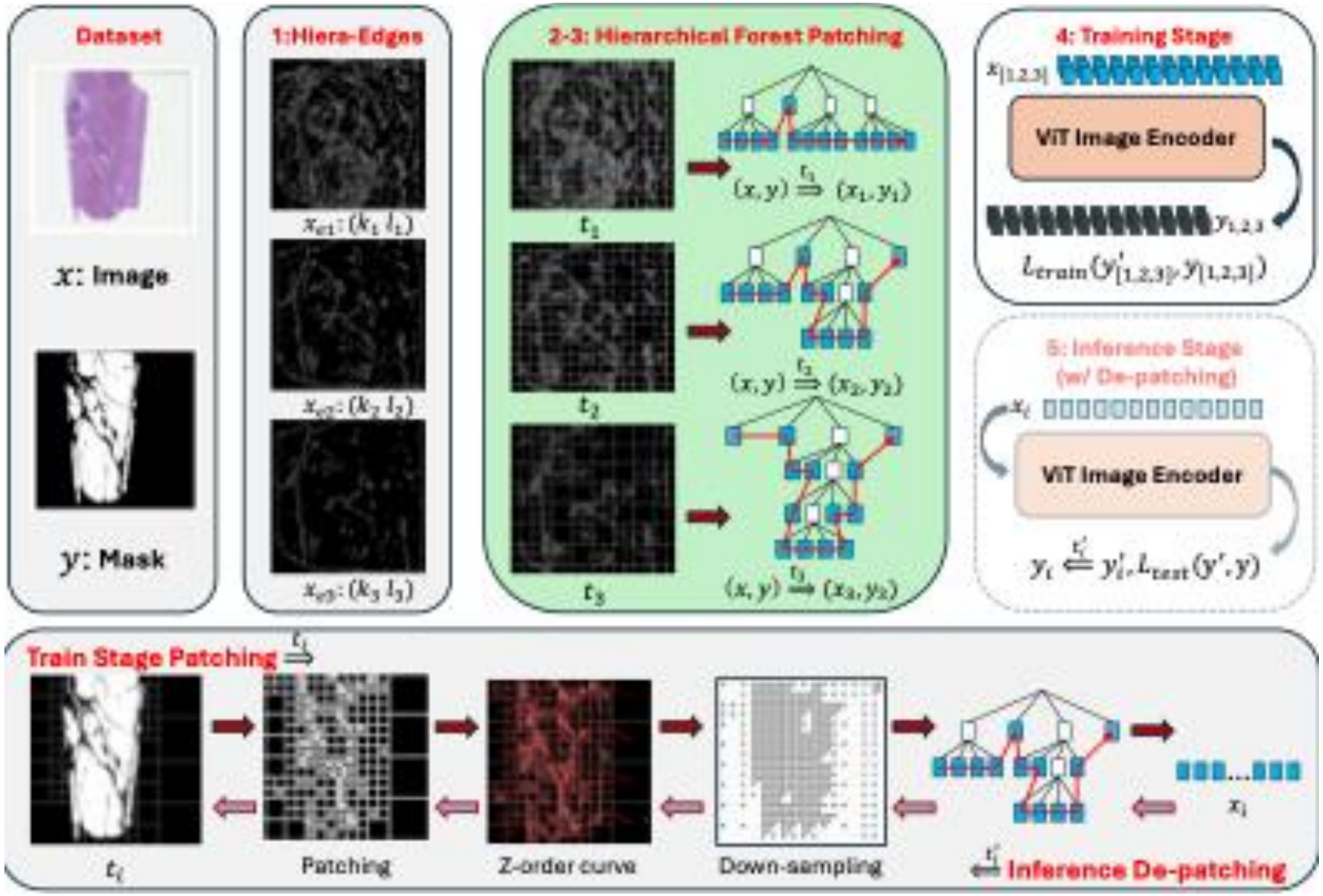
* https://arxiv.org/pdf/2408.00714

Meta's SAM 2 Model*

Symmetrical Hierarchical Forest with Pretrained ViT Encoder for High-Resolution Medical Segmentation

Remove Decoder

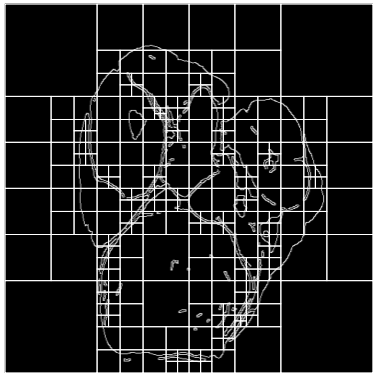
Apply a reverse depatching scheme to the output embeddings of the transformer encoder, eliminating the need for convolution-based decoders



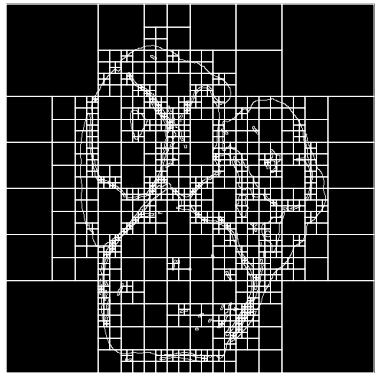
* E. Zhang et al. Under review

Symmetrical Hierarchical Forest with Pretrained ViT Encoder for High-Resolution Medical Segmentation

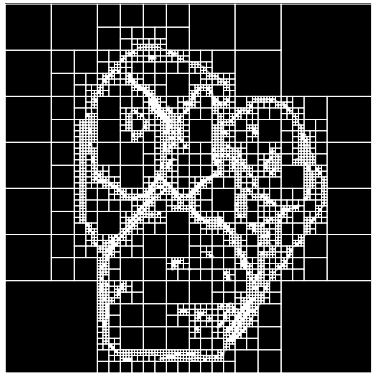
Resolution	Model	Patch	GPUs	Sec/Image/GPU	Depth	Sequence Length	Dice Score	Dice Improvement
512×512	**SAP+SAM**	8	1	0.0438	6	512	**75.31**	**+0.14**
	SAM	16	1	0.1983	-	1,024	61.39	
	AP+UNTER	8	1	0.0581	6	576	75.17	
	UNETR	16	1	0.1477	-	1,024	74.88	
	TransUNet	-	1	0.1783	-	-	73.32	
	UNet	-	1	0.0438	-	-	70.32	
$1,024 \times 1,024$	**SAP+SAM**	2	1	0.0991	9	1,024	**78.67**	**+0.25**
	SAM	8	8	0.3826	-	16,384	66.56	
	AP+UNTER	2	8	0.2314	9	1,024	78.42	
	UNETR	8	32	1.0863	-	16,384	75.72	
	TransUNet	-	8	1.3247	-	-	72.38	
	UNet	-	1	0.0981	-	-	68.92	
$4,096 \times 4,096$	**SAP+SAM**	2	8	0.3766	11	4,096	78.67	**+0.14**
	SAM	32	64	1.6183	-	16,384	71.05	
	AP+UNTER	2	128	0.6938	11	4,096	**79.63**	
	UNETR	32	128	1.8613	-	16,384	75.77	
	TransUNet	-	128	2.1637	-	-	71.32	
	UNet	-	16	0.3712	-	-	64.11	
$8,192 \times 8,192$	**SAP+SAM**	2	16	1.5327	12	8192	**79.68**	**+0.12**
	SAM	64	128	2.5168	-	16,384	67.31	
	AP+UNTER	2	256	2.3314	12	10,609	79.56	
	UNETR	64	256	2.6618	-	16,384	75.27	
	TransUNet	-	256	2.3678	-	-	70.89	
	UNet	-	32	1.2858	-	-	63.21	
$16,384 \times 16,384$	**SAP+SAM**	2	32	3.2741	13	16,384	**80.98**	**+0.36**
	SAM	128	256	5.6714	-	16,384	67.63	
	AP+UNTER	2	512	4.8792	13	16,384	80.62	
	UNETR	128	512	5.1179	-	16,384	75.89	
	TransUNet	-	512	6.1296	-	-	70.46	
	UNet	-	256	2.7825	-	-	62.97	
$32,768 \times 32,768$	**SAP+SAM**	4	64	3.4631	13	16,384	**81.43**	**+2.45**
	SAM	256	512	9.1213	-	16,384	62.34	
	AP+UNTER	4	1024	7.8916	13	16,384	78.98	
	UNETR	256	1024	8.1896	-	16,384	74.96	
	TransUNet	-	1024	10.001	-	-	69.88	
	UNet	-	512	4.2714	-	-	61.38	
$65,536 \times 65,536$	**SAP+SAM**	8	256	3.6112	13	16,384	**82.96**	**+5.19**
	SAM	1024	1024	12.983	-	16,384	61.68	
	AP+UNTER	8	2048	12.697	13	16,384	77.77	
	UNETR	1024	2048	13.218	-	16,384	75.31	
	TransUNet	-	2048	14.352	-	-	67.67	
	UNet	-	1024	5.961	-	-	59.69	



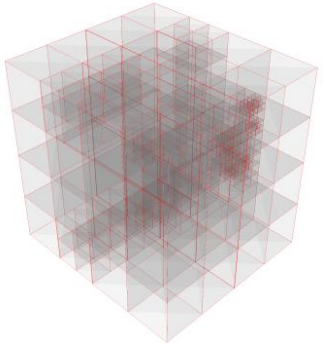
(b) Sequence Length=256



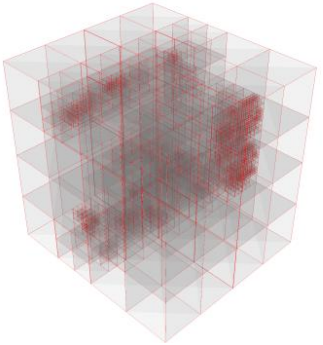
(c) Sequence Length=1024



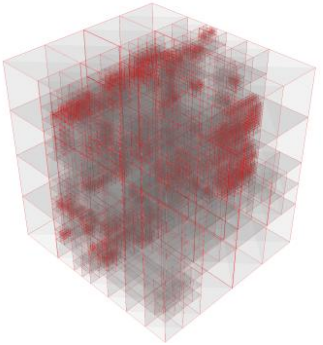
(d) Sequence Length=4096



(f) Sequence Length=260



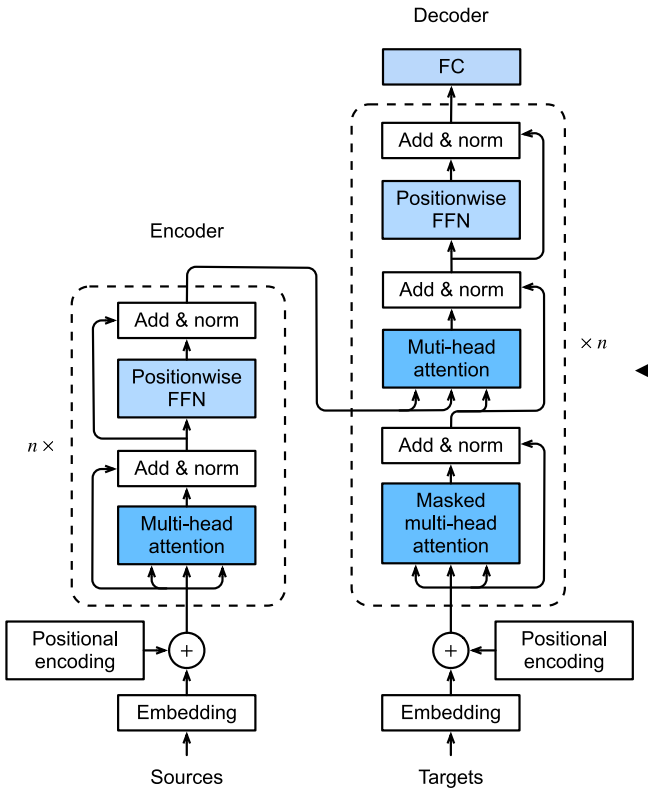
(g) Sequence Length=1030

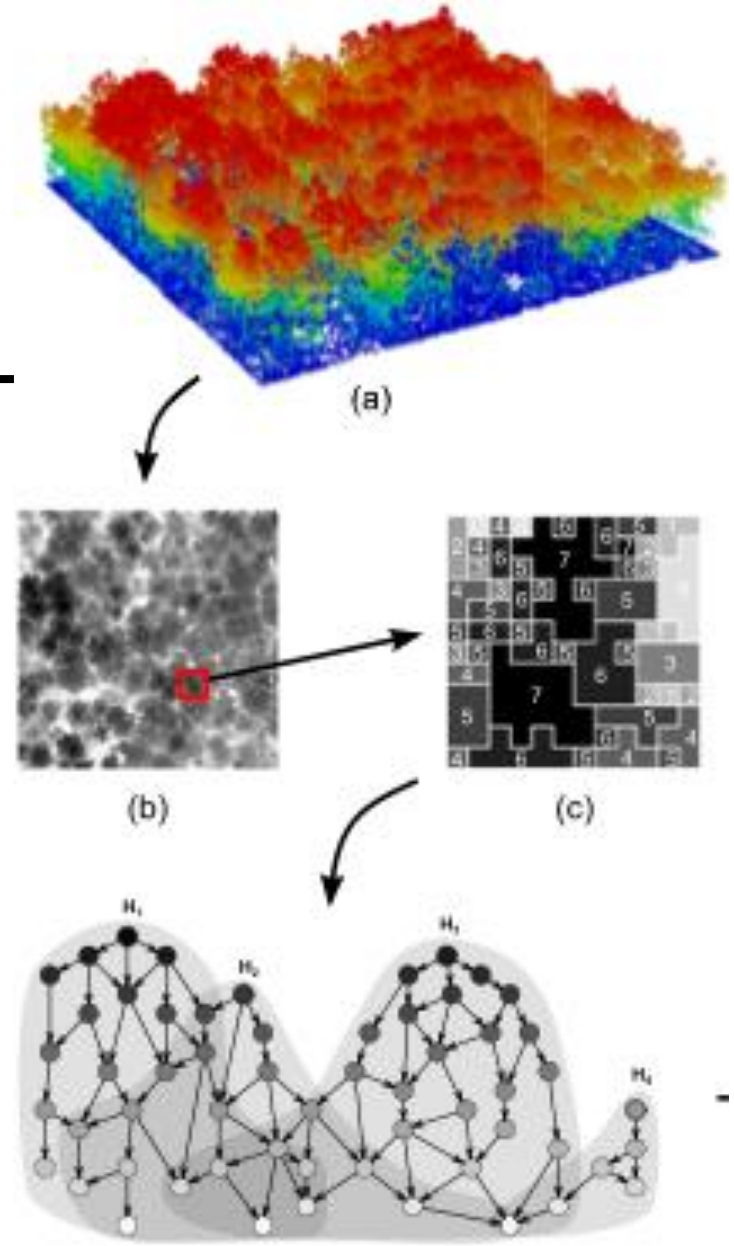


(h) Sequence Length=4096

Can ViT Learn Where to Look?

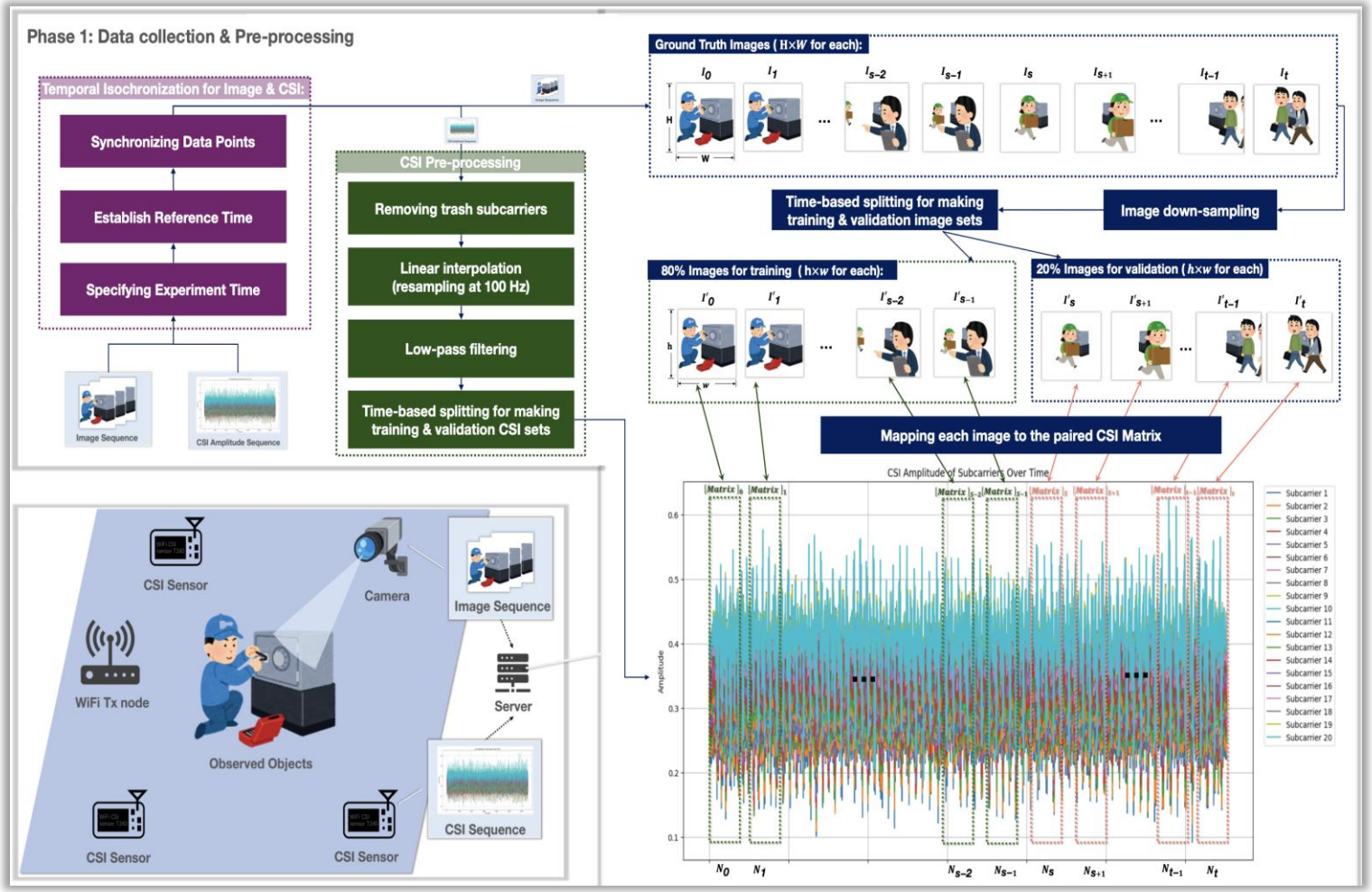
- HUMANS tell the model where to look
- Can the model learn where to look?
  - by being fed the spatial hierarchy

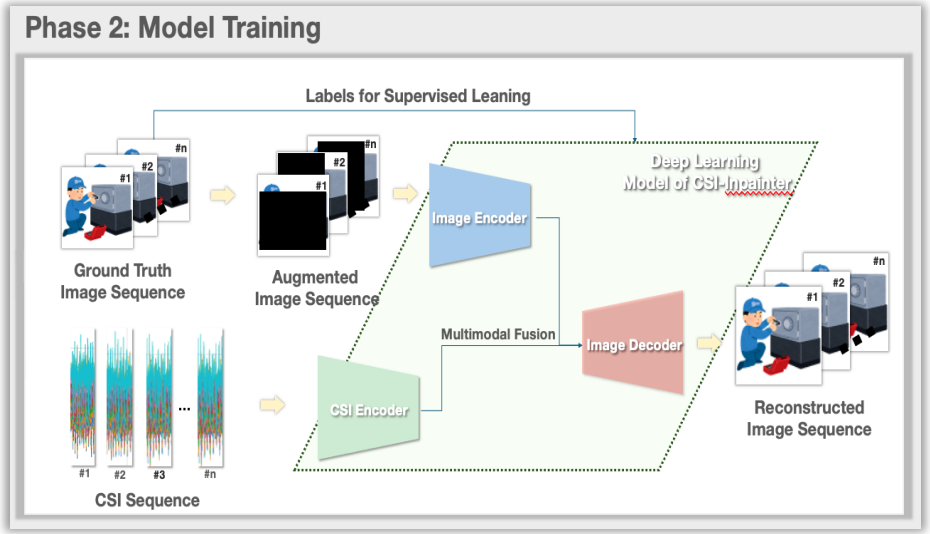


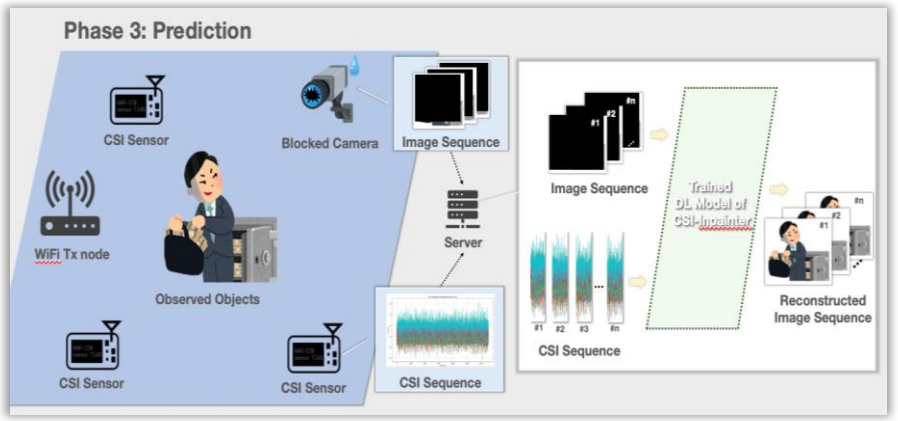


Visual Scene Recovery from Wi-Fi CSI

➢CSI-Inpainter: CSI-guided obstacle removal

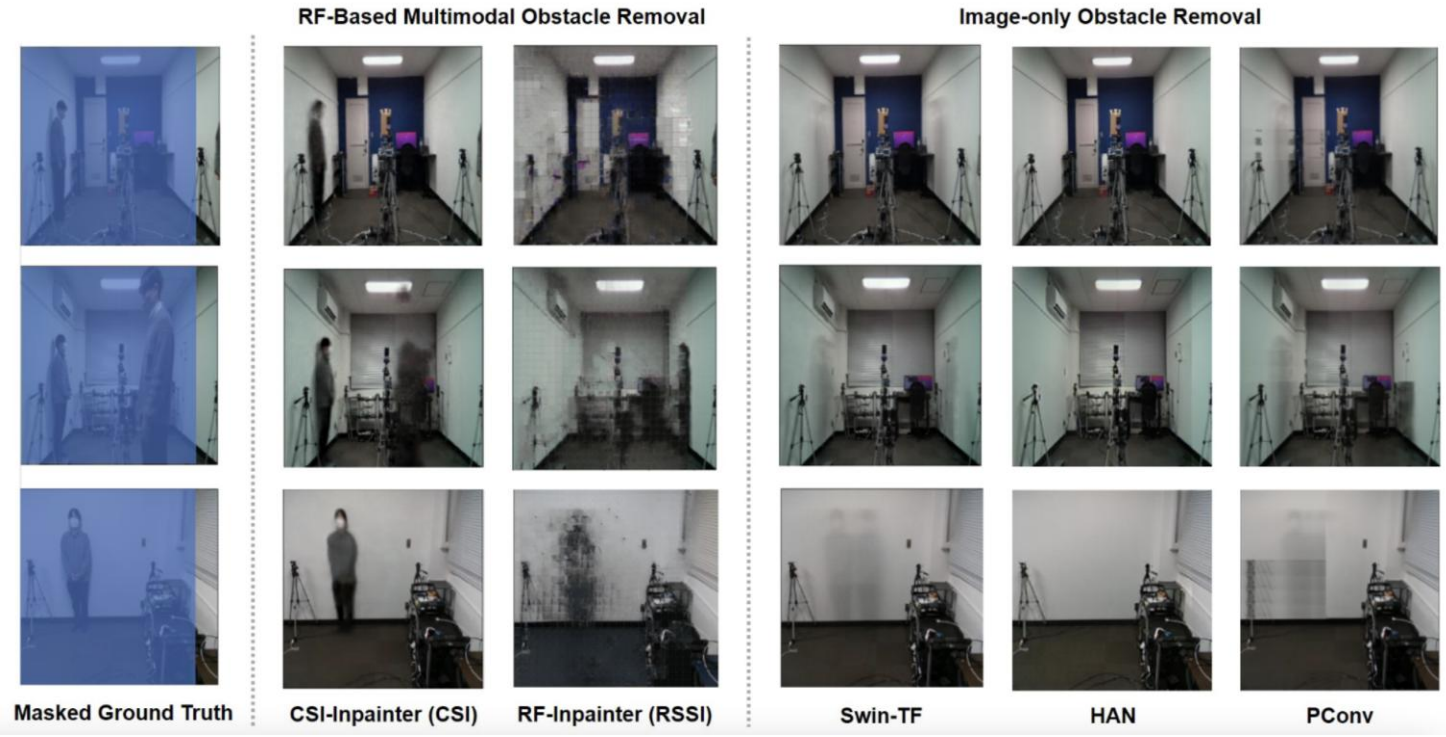


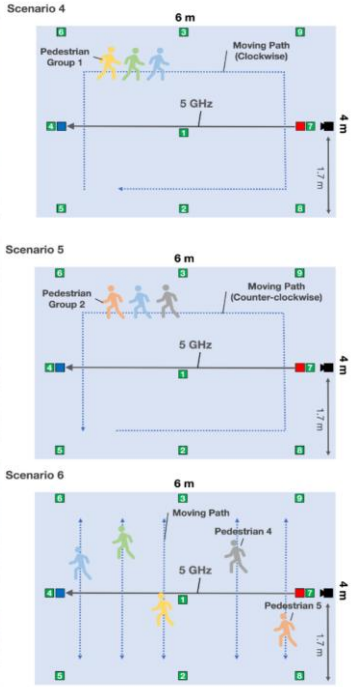




* Chen et al, Trans-Inpainter: A Transformer Model for High Accuracy Image Inpainting from Channel State Information, IEEE IoT'25

Visual Scene Recovery from Wi-Fi CSI

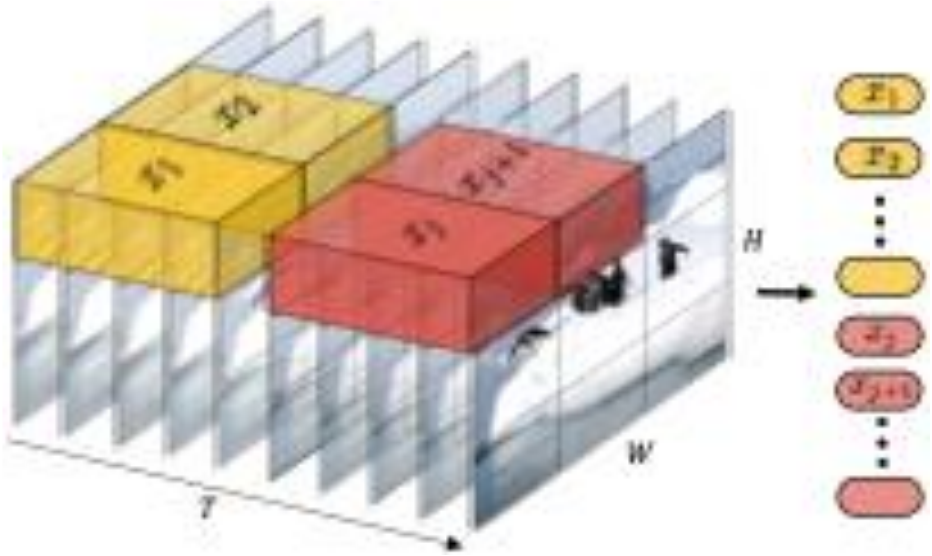




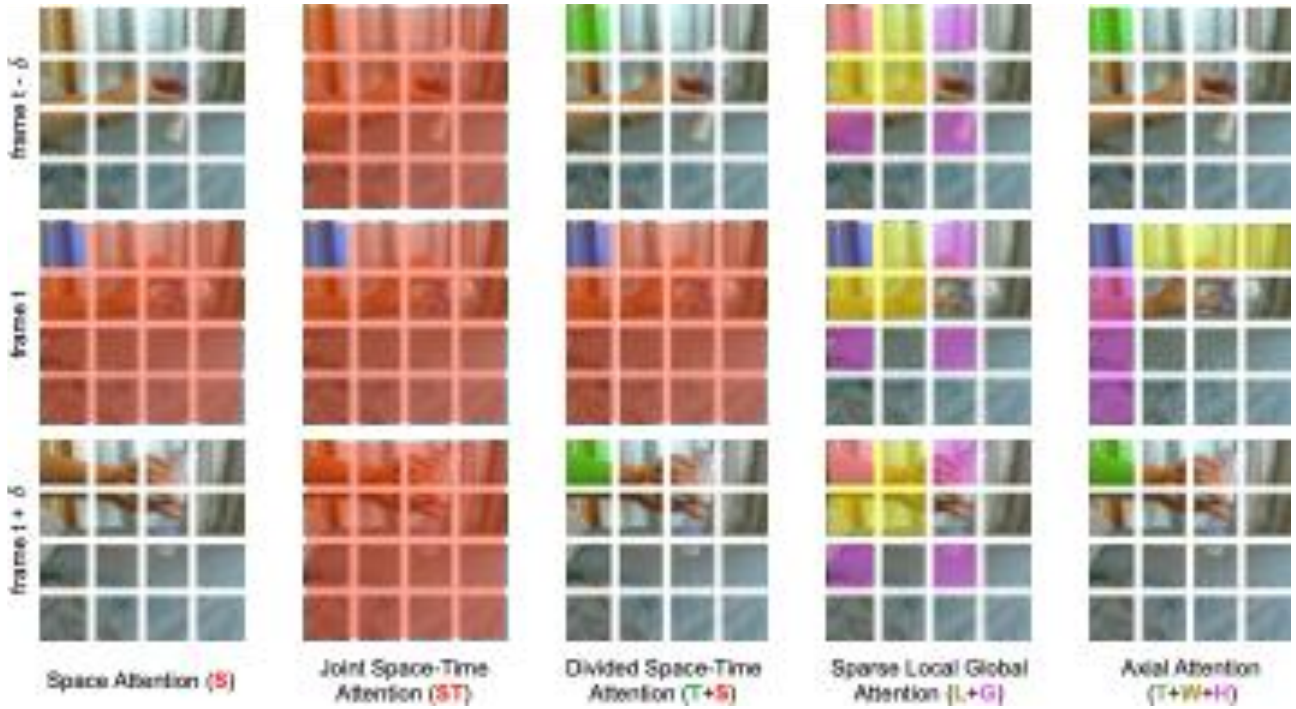
ViT Challenges

1. Long sequence (when ingesting all image elements at high res)
2. Shifting bottleneck (elements outside the Transformer encoder)
3. Tokenization (managing temporal dimension)
4. Different parallelism in training (vs. text Transformer)
5. Multi-modality

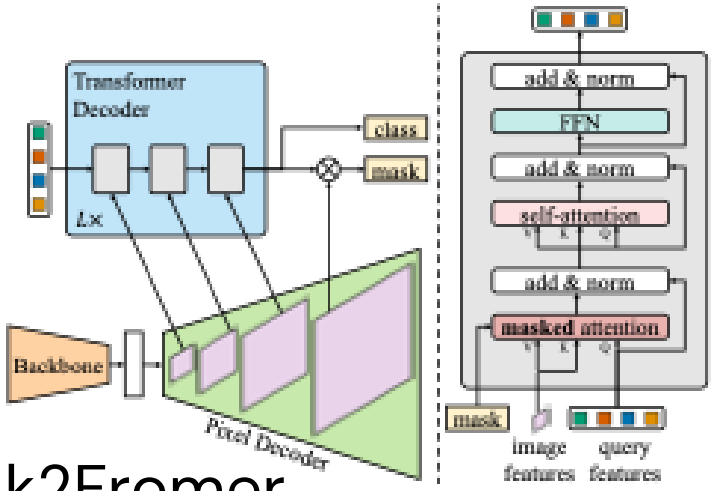
Tokenizing Spatio-temporal Data is Tricky



ViViT



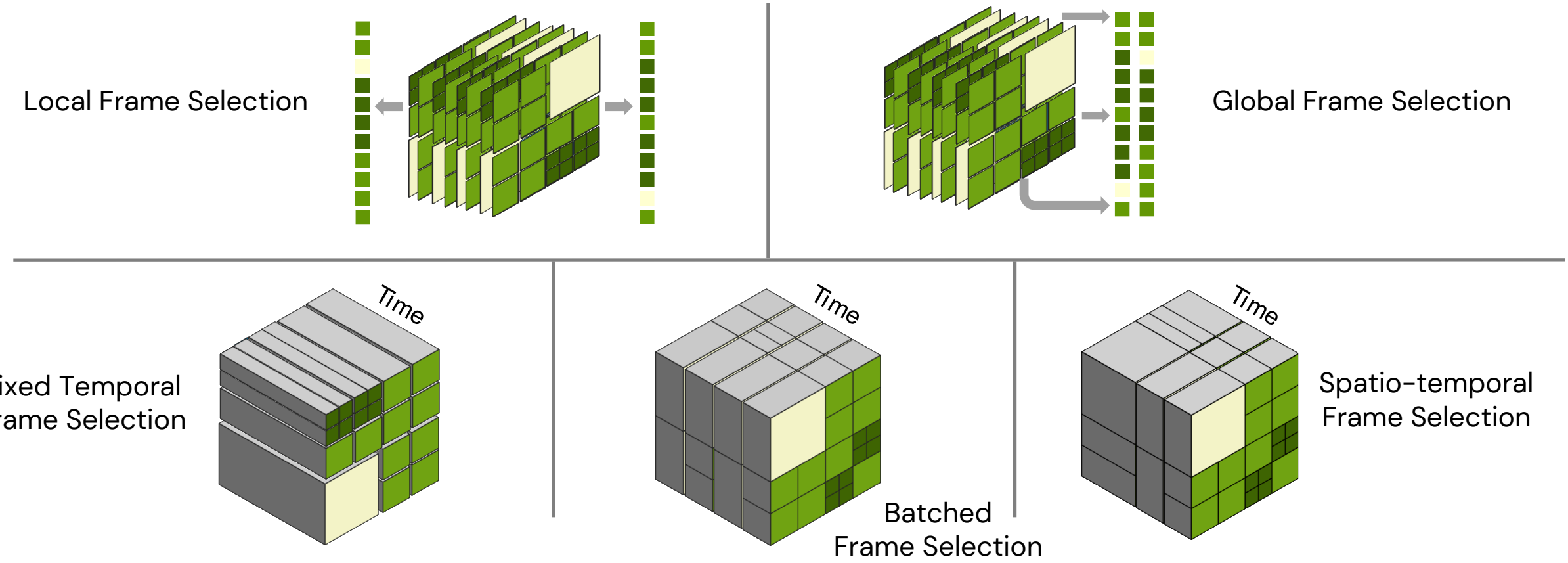
TimesFormer



Mask2Fromer

How to Tokenize Spatio-temporal Data

➢Different tokenization schemes aligned with adaptive pathcing

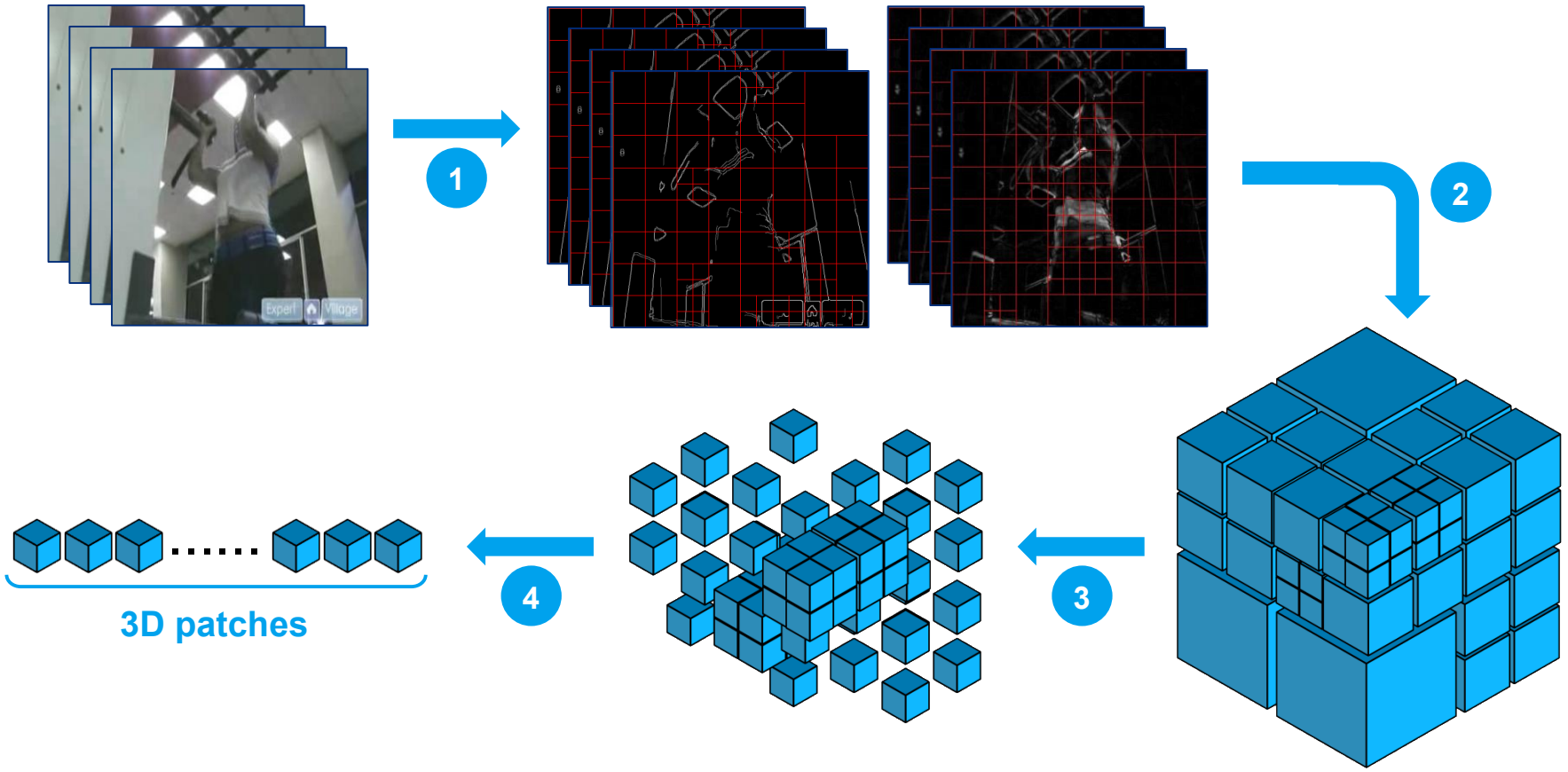


We observe how Local and Global Frame Selection operate on a frame-by-frame basis to form quadtrees, while the other three schemes work across three dimensions. Among these three, Batch and Temporal-Spatial Frame Selection divide the time dimension, with the former using even divisions and the latter using uneven divisions.

Adaptive Patching for Videos

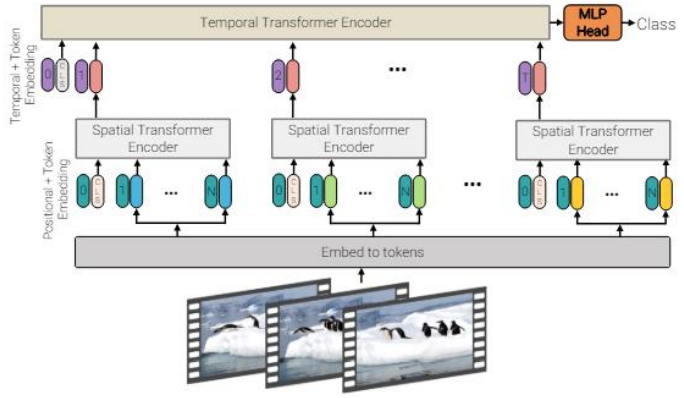
- For videos we adaptively patch temporally
  - merging spatial and temporal adaptiveness?

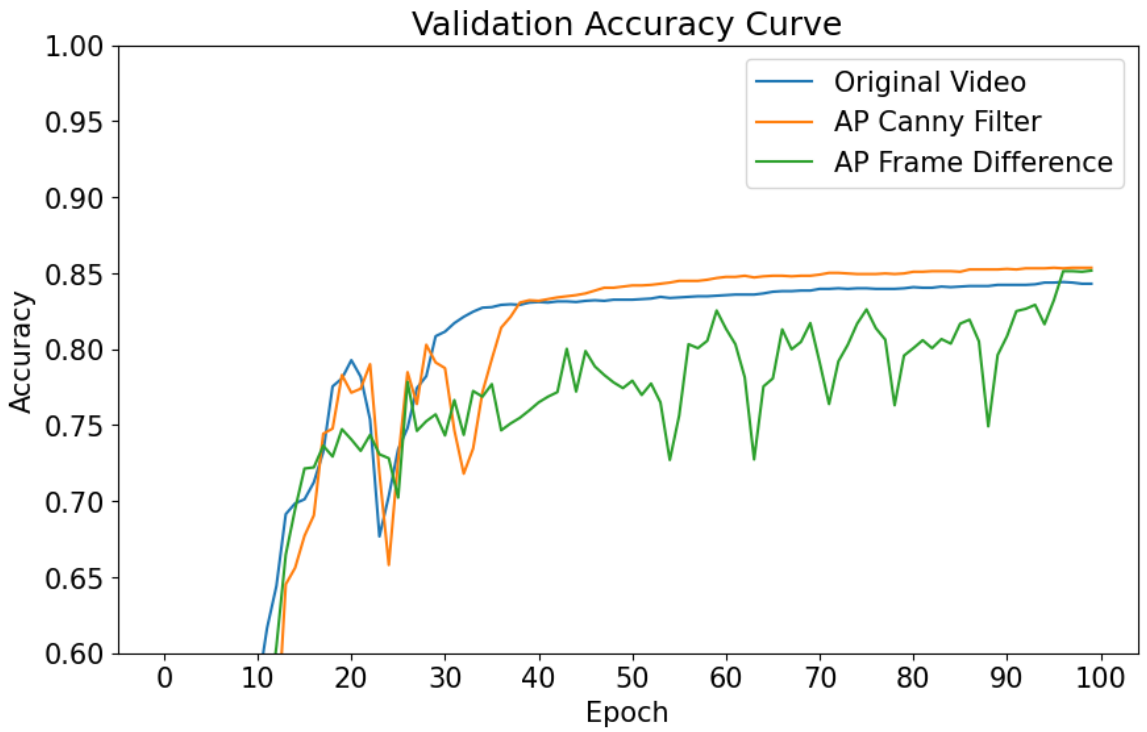
    customize the adaptive scheme based on the nature of the input dataset? AP as a means for compression, followed by a scheme to arrange patches to recreate the input video?

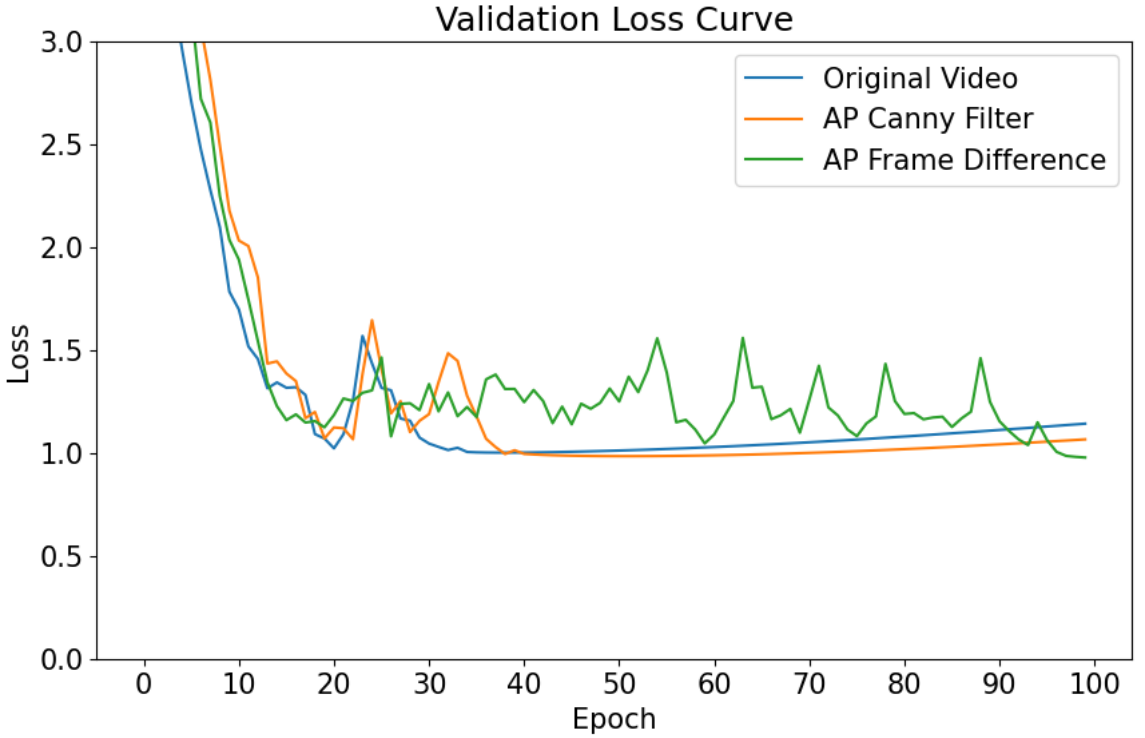


Application in Video Action Recognition

- We use the **Video Vision Transformer (ViViT)** model for this task, containing attention in the spatial and temporal dimensions
- AP able to achieve comparable metrics with **up to 4x memory reduction** while maintaining the same number of patches

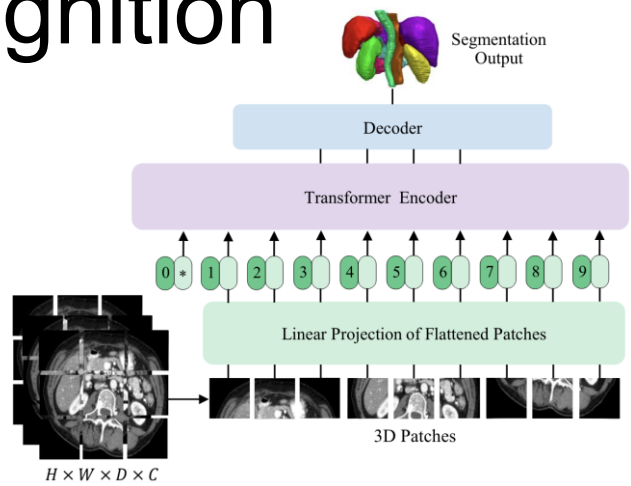


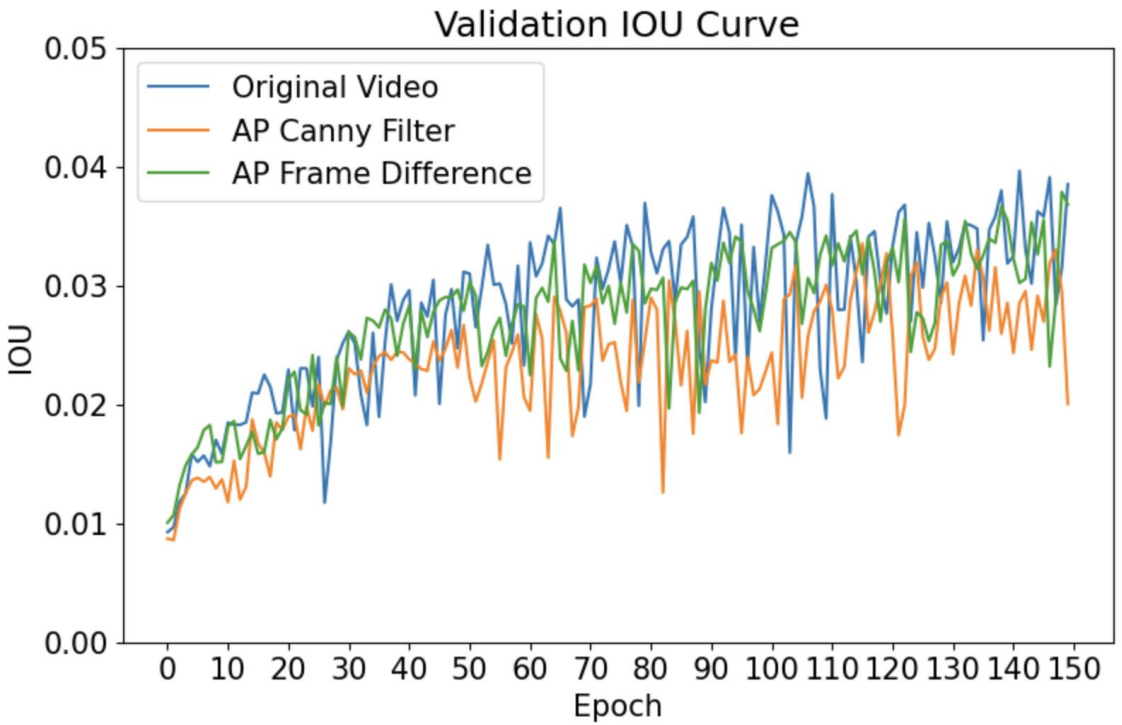


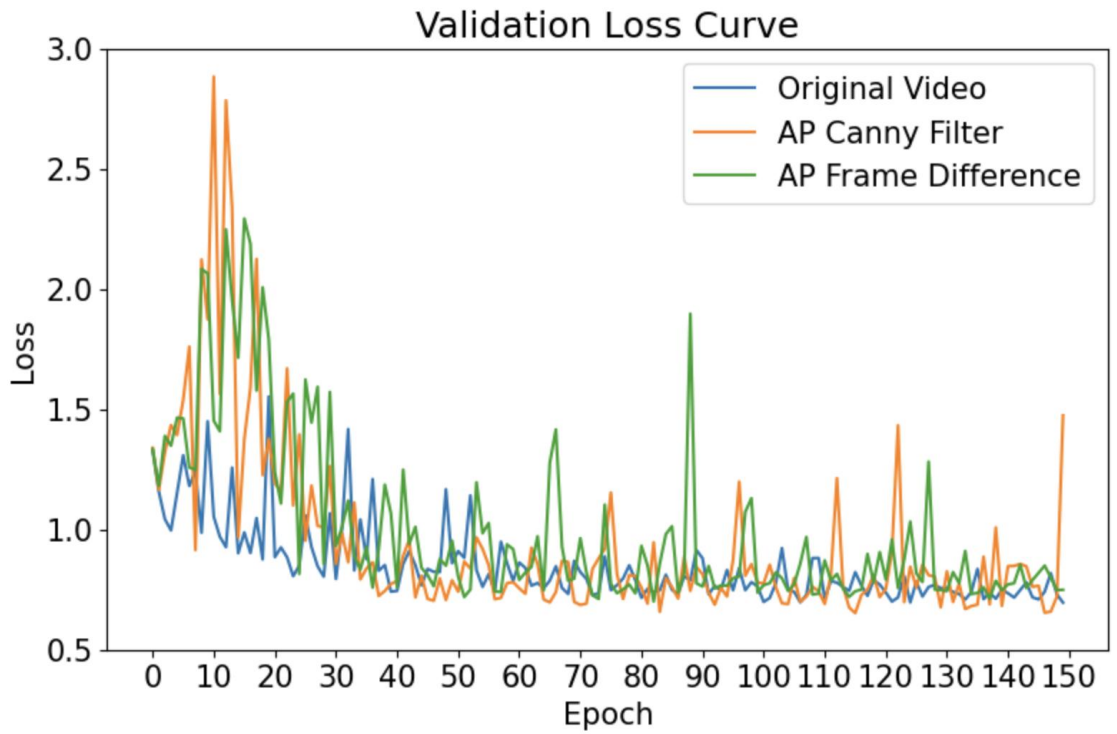


Application in Video Action Recognition

- We use the **UNEt TRansformers (UNETR)** model for this task, combining a transformer encoder with a convolutional decoder
- AP able to achieve comparable metrics with **up to 8x memory reduction** while maintaining the same number of patches



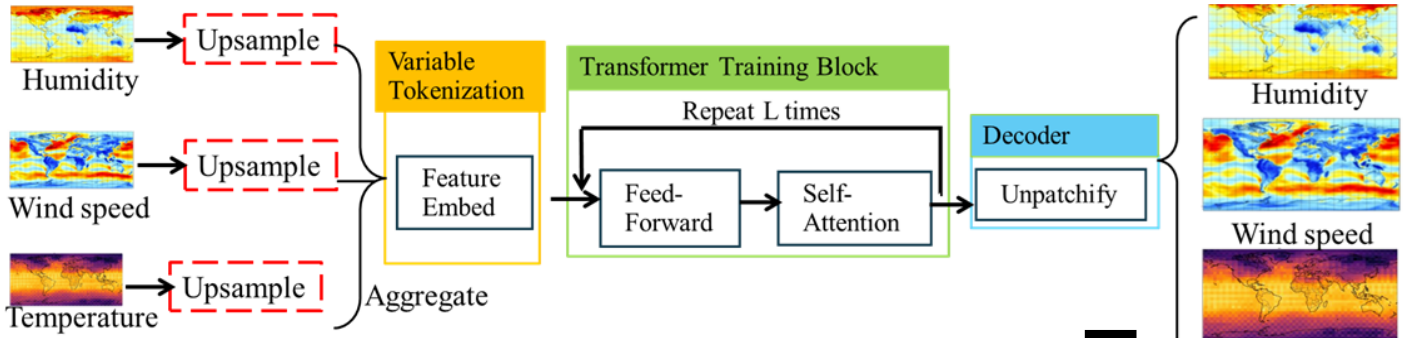




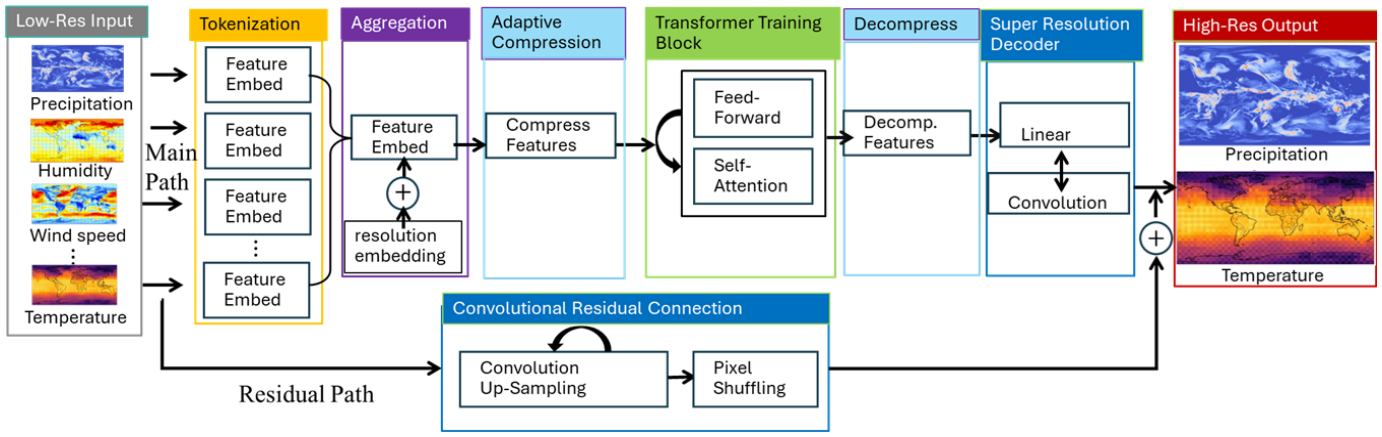
ViT Challenges

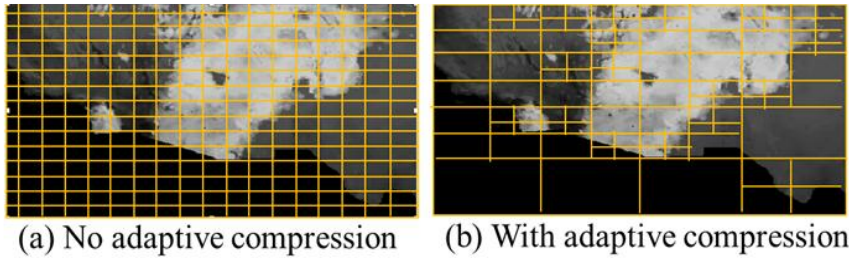
1. Long sequence (when ingesting all image elements at high res)
2. Shifting bottleneck (elements outside the Transformer encoder)
3. Tokenization (managing temporal dimension)
4. Different parallelism in training (vs. text Transformer)
5. Multi-modality

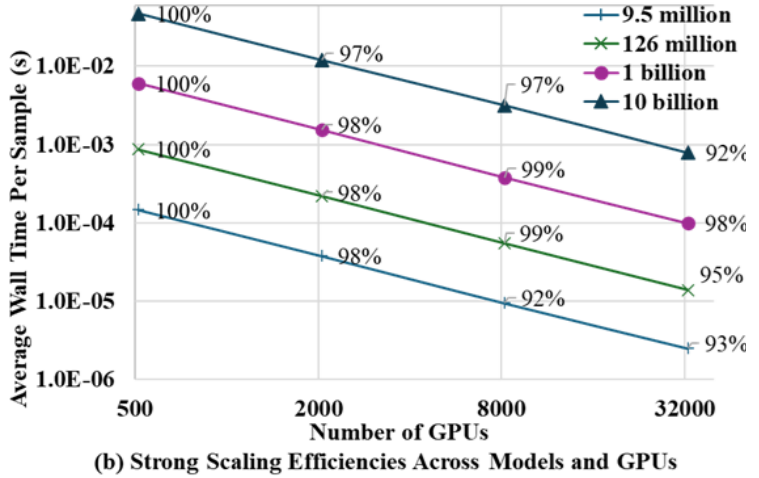
ORBIT-2: Scaling Exascale Vision Foundation Models for Weather and Climate Downscaling



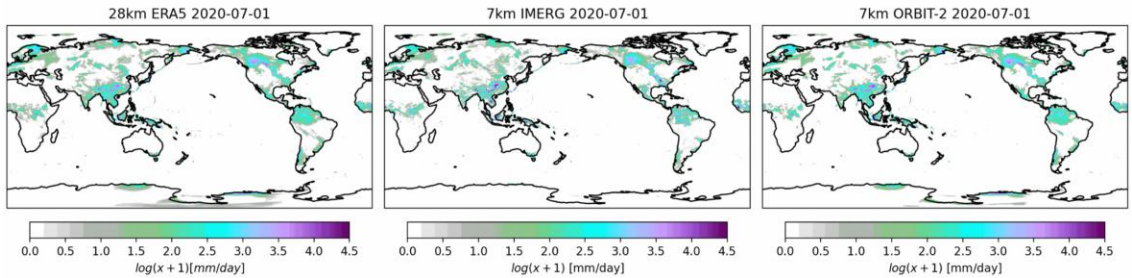
- Operating directly on compressed inputs
- Condition prediction for reduced uncertainty







(b) Strong Scaling Efficiencies Across Models and GPUs

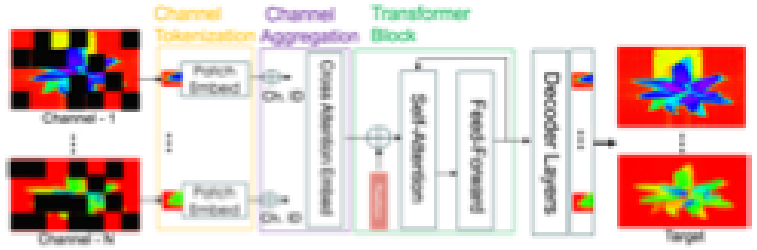


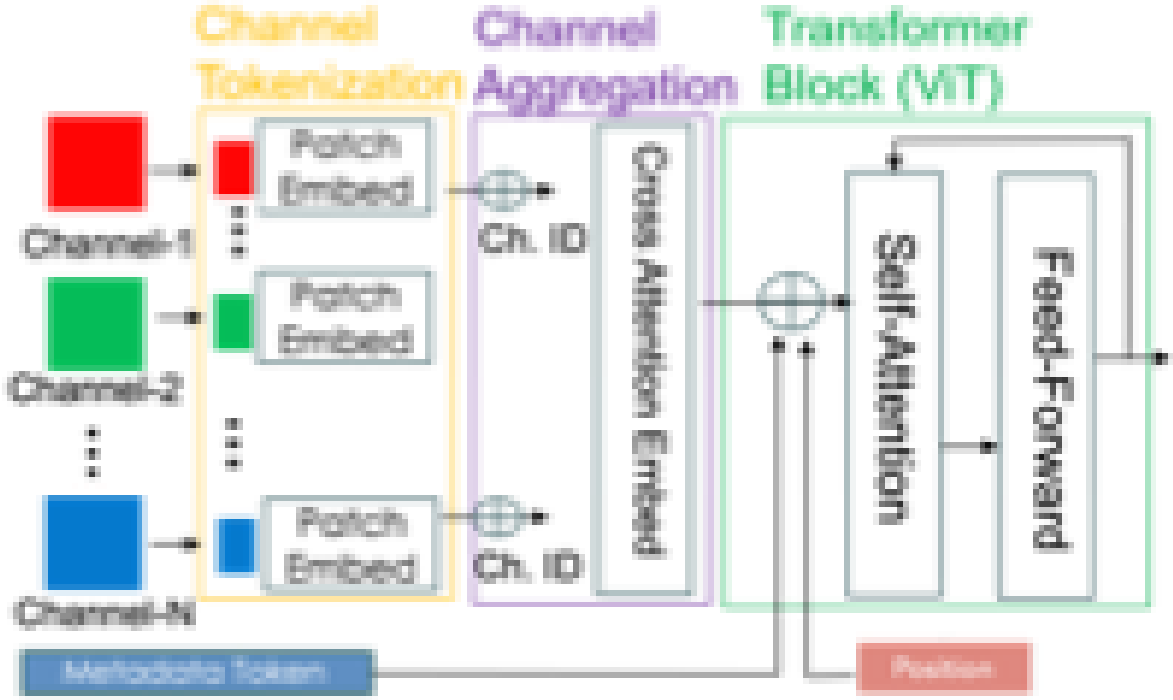
* Wang et al. To appear SC'25 (Gordon Bell Finalist)

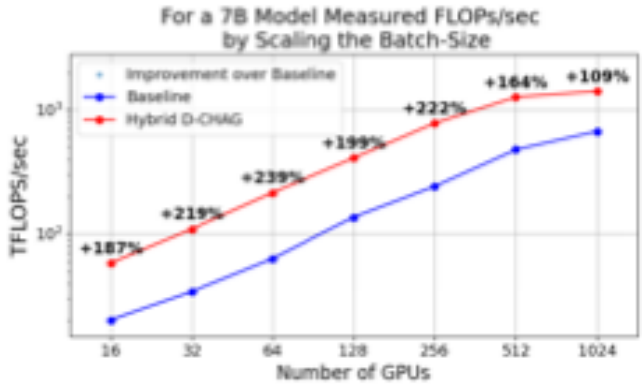
Distributed Cross-Channel Hierarchical Aggregation for Foundation Models

Distributed Tokenization

For high number of channels: distribute tokenization and implement a hierarchical strategy for channel aggregation.





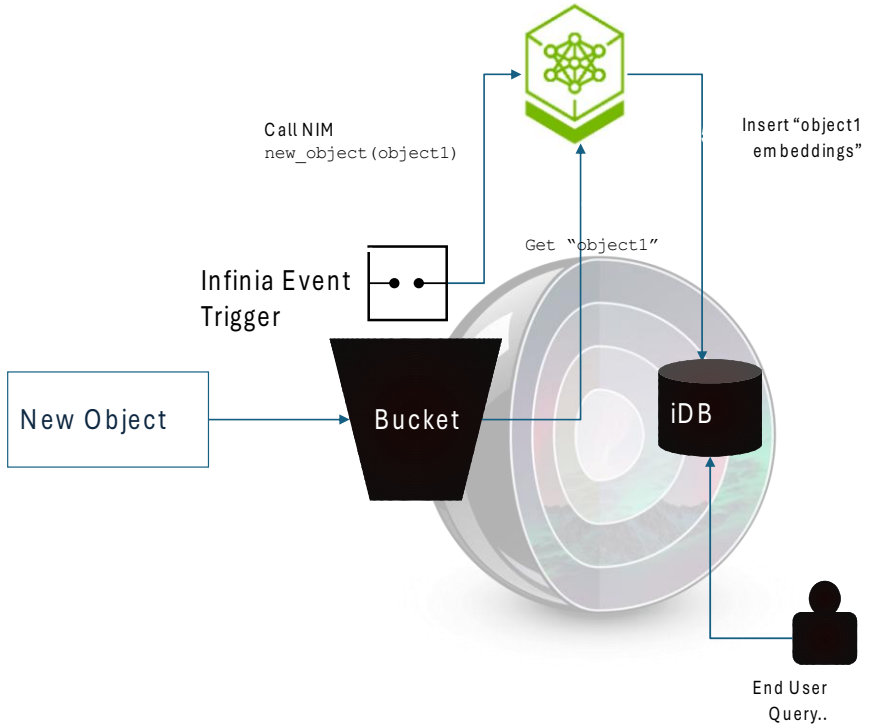


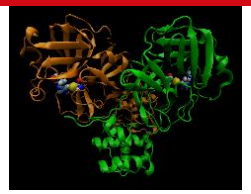
* Tsaris et al. To appear SC'25

ViT Challenges

1. Long sequence (when ingesting all image elements at high res)
2. Shifting bottleneck (elements outside the Transformer encoder)
3. Tokenization (managing temporal dimension)
4. Different parallelism in training (vs. text Transformer)
5. Multi-modality

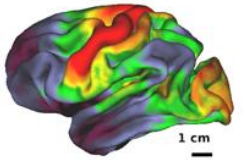
Multi-agent Approach for Multi-modality





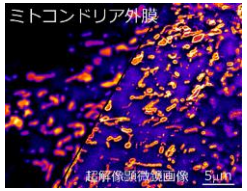
Proteins

Information Retrieval APIs (or Code)



Neural Activities

Information Retrieval APIs (or Code)

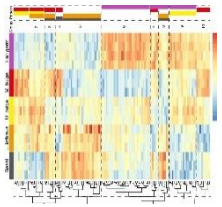


Images

Information Retrieval APIs (or Code)

Omics lipidome

Information Retrieval APIs (or Code)



Genome/Transcriptome

Information Retrieval APIs (or Code)

Other Phenotypes

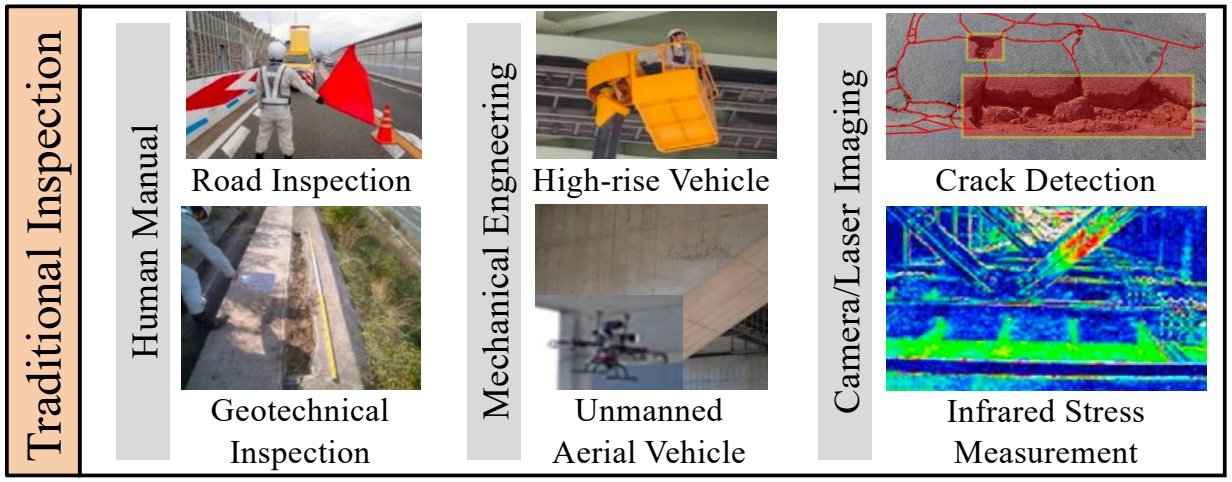
Information Retrieval APIs (or Code)

Real-world Problem Covering Almost all the Challenges Presented so Far

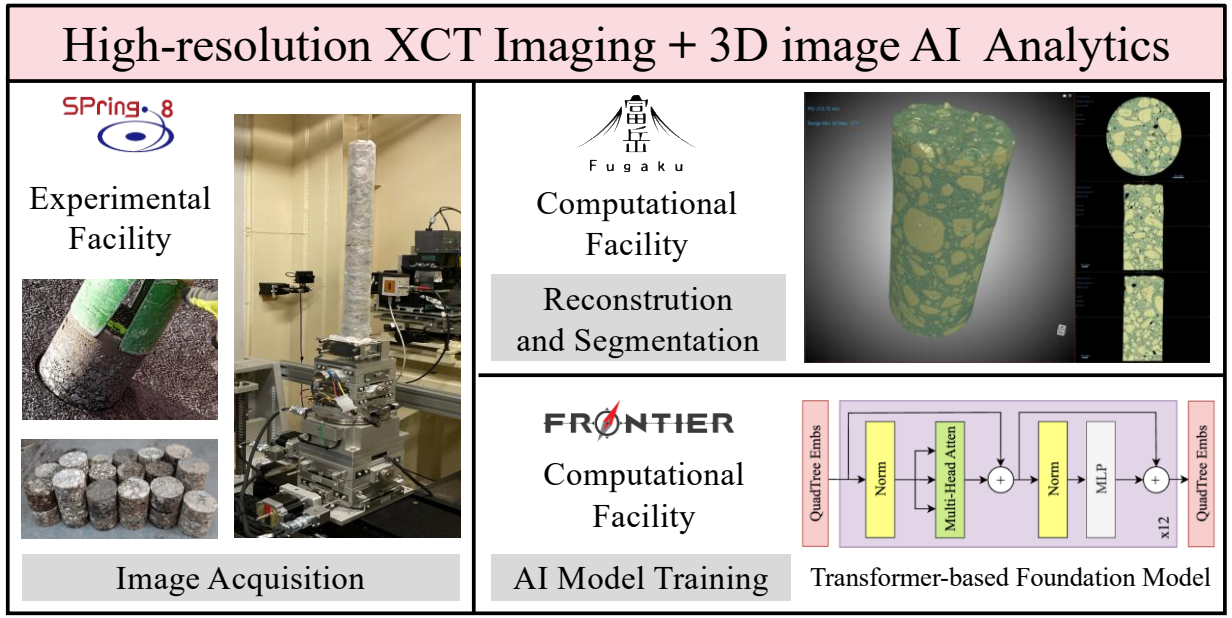
Paradigm Shift in Infrastructure Inspection Technology

High performance Imaging + AI + analytics

- Replace traditional engineering
- Full fusion of imaging and AI



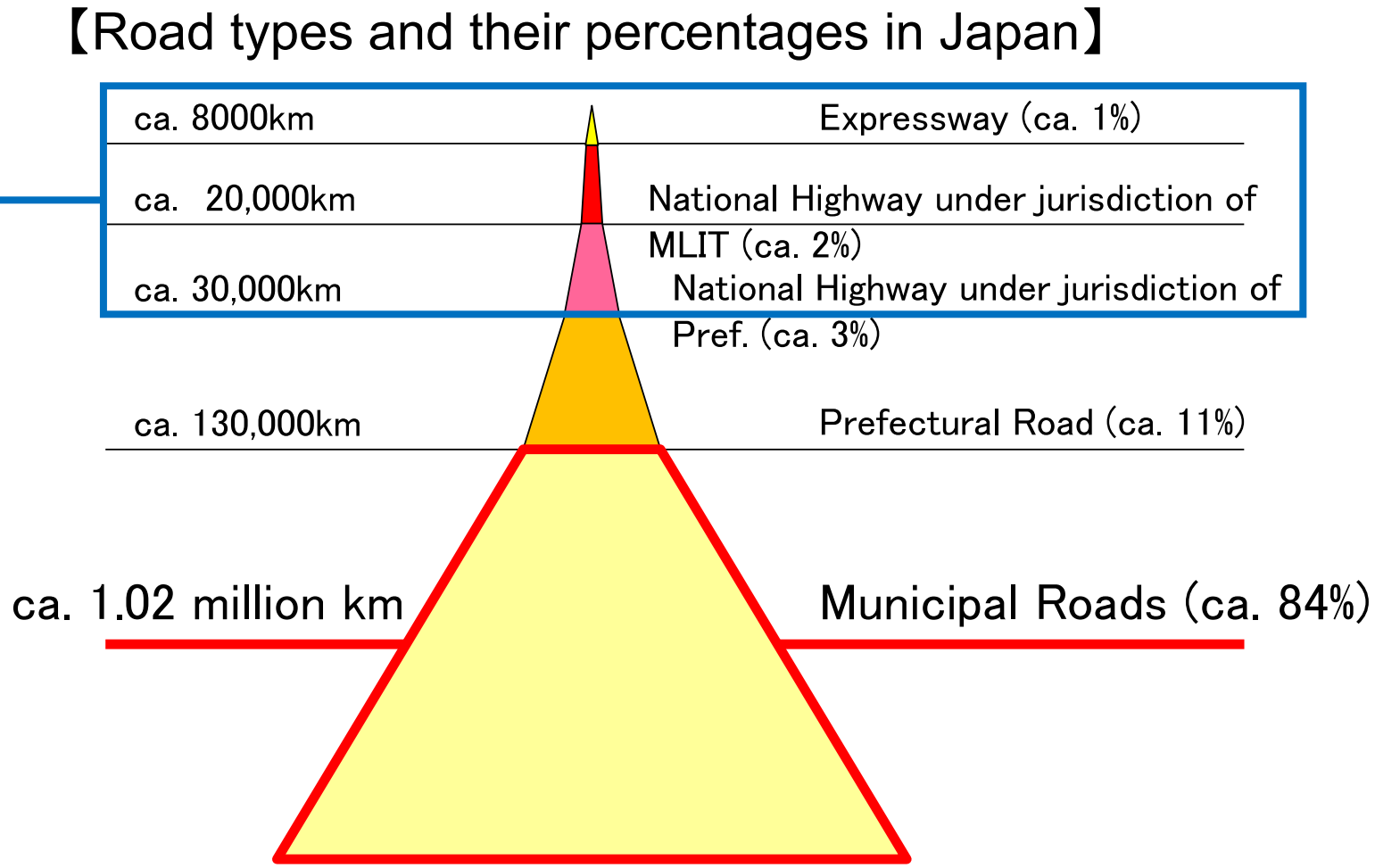
Infrastructure Inspection Paradigm Shift



https://tinyurl.com/yvddayb3

Total Road Length in Japan is ca. 1.21 Million KMs

Total 130,000km of highway to maintain



* *Source: MLIT, Japan https://www.mlit.go.jp/road/road_e/pdf/RoadMaintenance.pdf*

Highway Network West Japan (Osaka, Kyoto, Kobe)

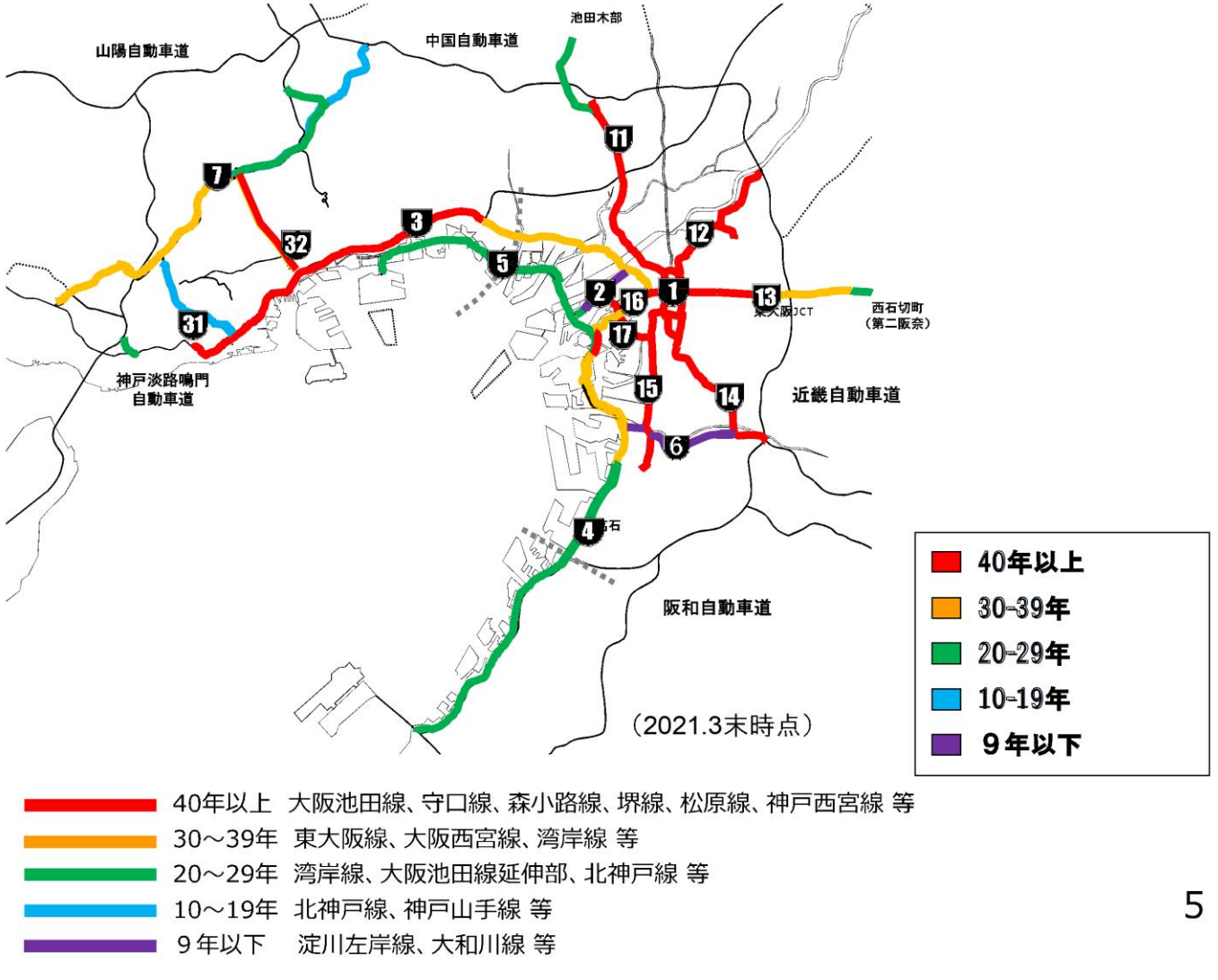
1号環状線　四ツ橋付近

1964年6月28日、土佐堀～湊町間が開通

大和川第三トンネル

（6号大和川線：供用中）

淀川左岸線（2期）建設中



40年以上　大阪池田線、守口線、森小路線、堺線、松原線、神戸西宮線 等

30～39年　東大阪線、大阪西宮線、湾岸線 等

20～29年　湾岸線、大阪池田線延伸部、北神戸線 等

10～19年　北神戸線、神戸山手線 等

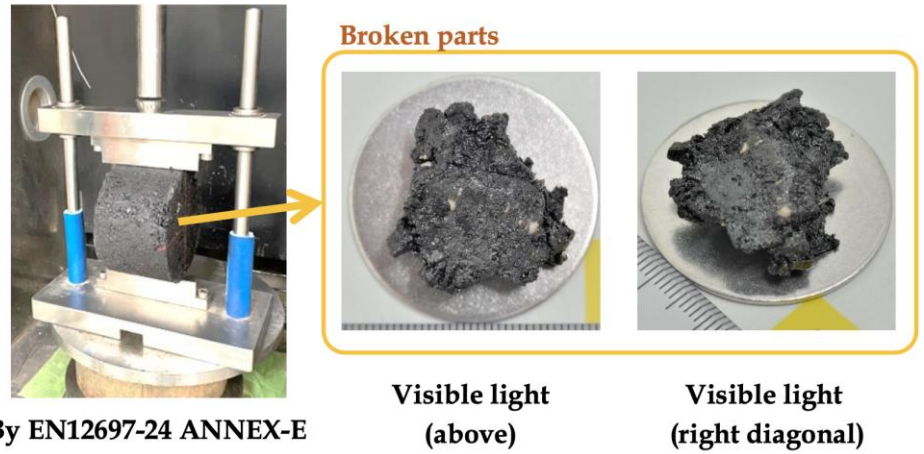
９年以下　淀川左岸線、大和川線 等

5

** Source: Sakai @Hanshin Highway*

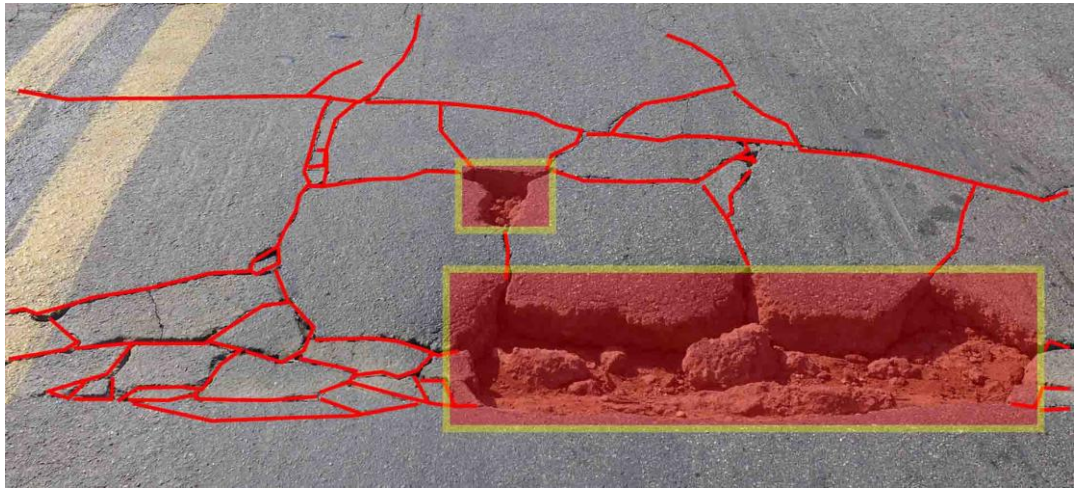
How to Inspect Roads for Maintenance?

- Mechanical inspection
  - Time: 10s of years
  - Cost: $ Billions

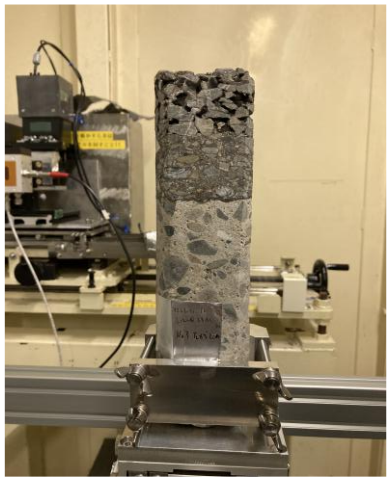


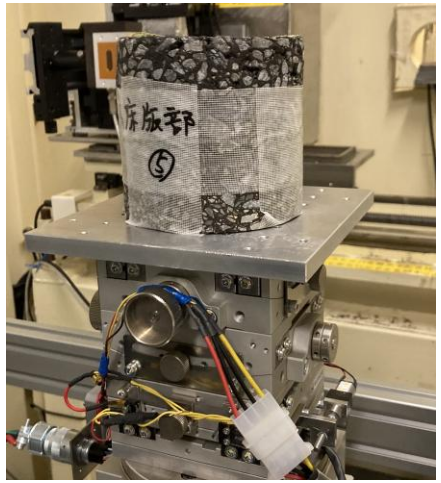
Fatigue test: Accelerated Crack Simulation

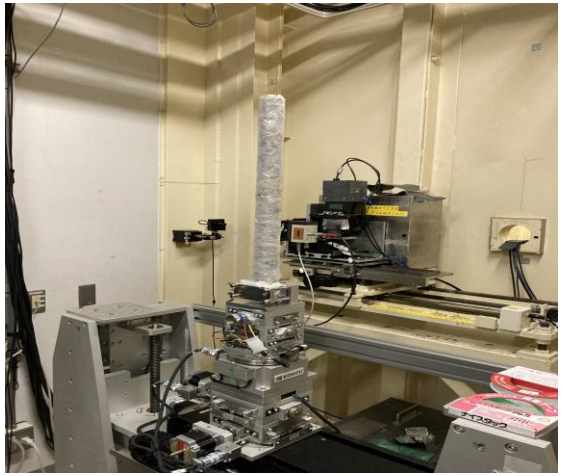
- Camera/laser Imaging technology
  - Good for fast screening of visible surface cracks, depressions etc
  - Not a reliable technology for understanding sub-surface conditions

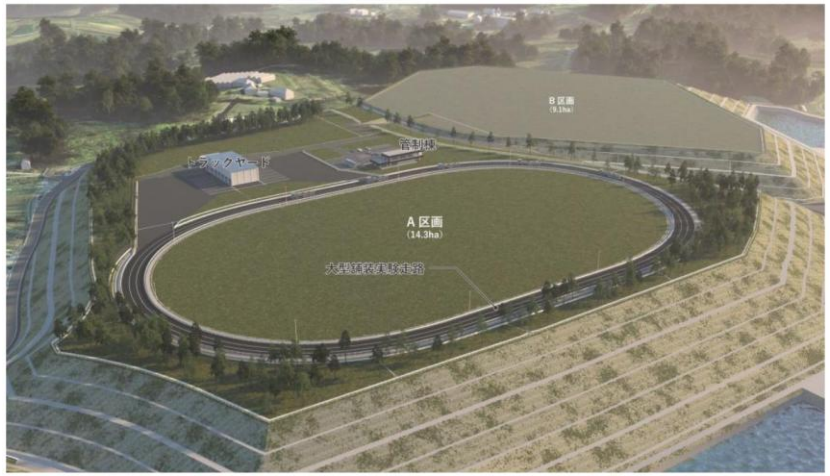


Mechanical Inspection









"Actual-scale test track" by Taisei-Rotech

Vehicles Extracting 100s Samples per Day

- Core samples extraction machines mounted on vehicles

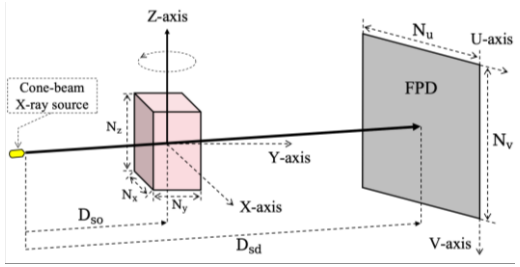
+

Collecting Samples

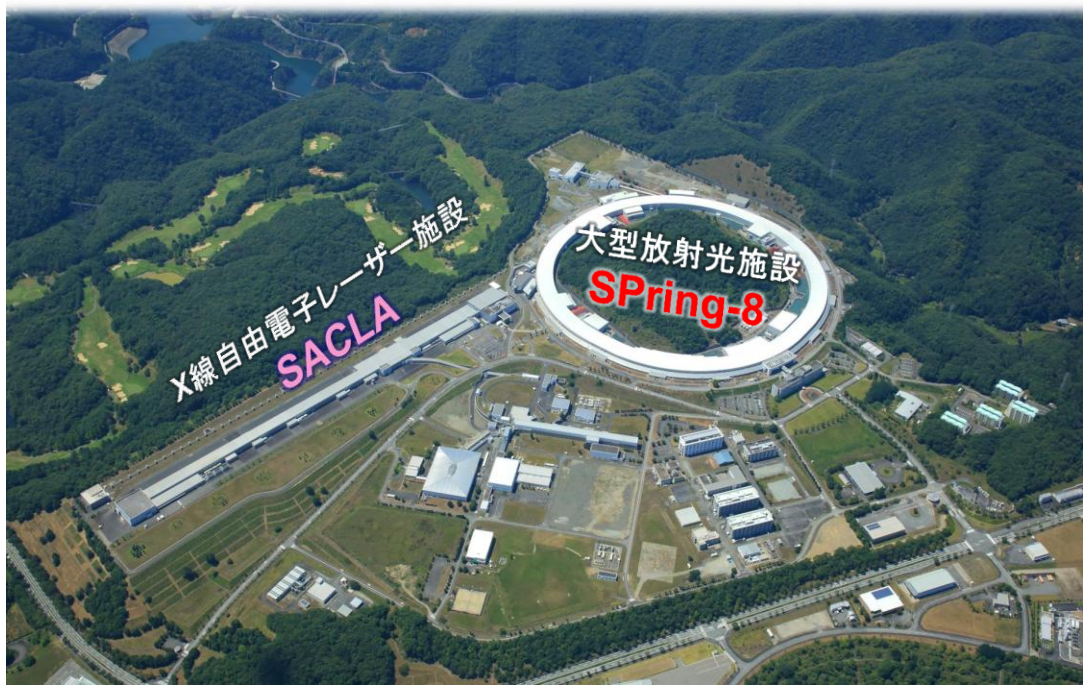
- Extract cylindrical samples from core of asphalt layers

** Example: if one sample every 10KM, then 130,000 / 10 =* ***13,000 samples***

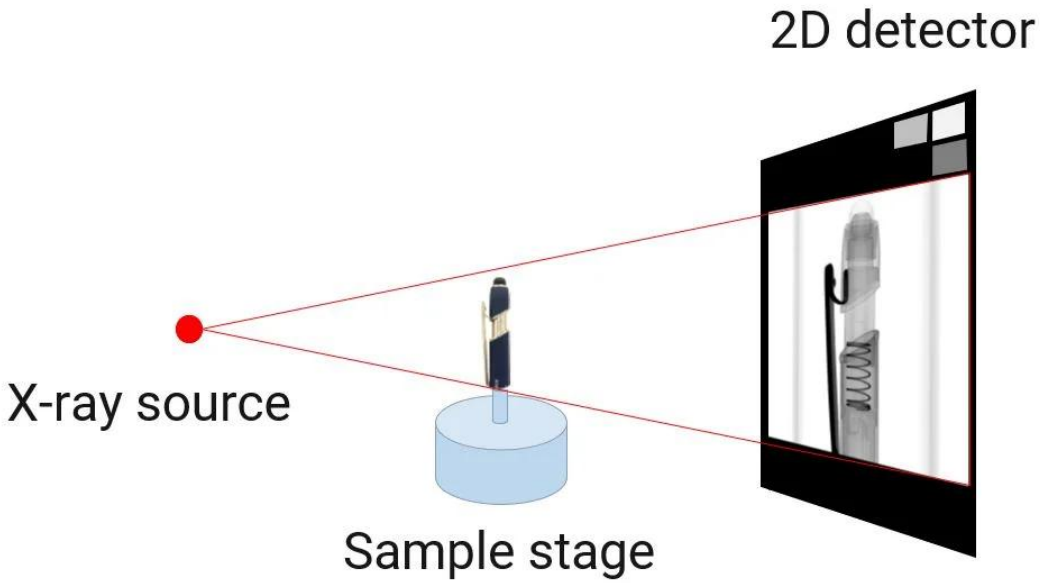
Scan Samples with X-ray



- Scan (2D projections) at RIKEN Spring-8 Synchrotron



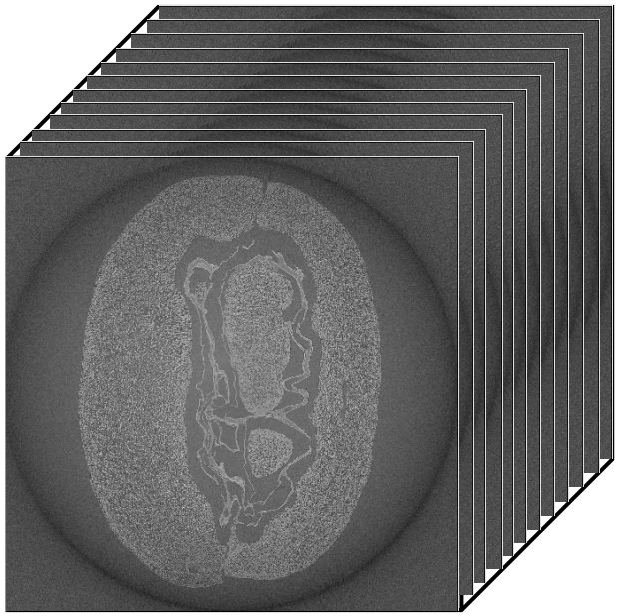
The synchrotron radiation facility RIKEN SPring-8



Move Data to RIKEN-CCS Fugaku

- Move projections data to R-CCS (Fugaku)

10,000s of Samples*



8192^3

2TB

Fastest way to transfer data is to ship HDD!

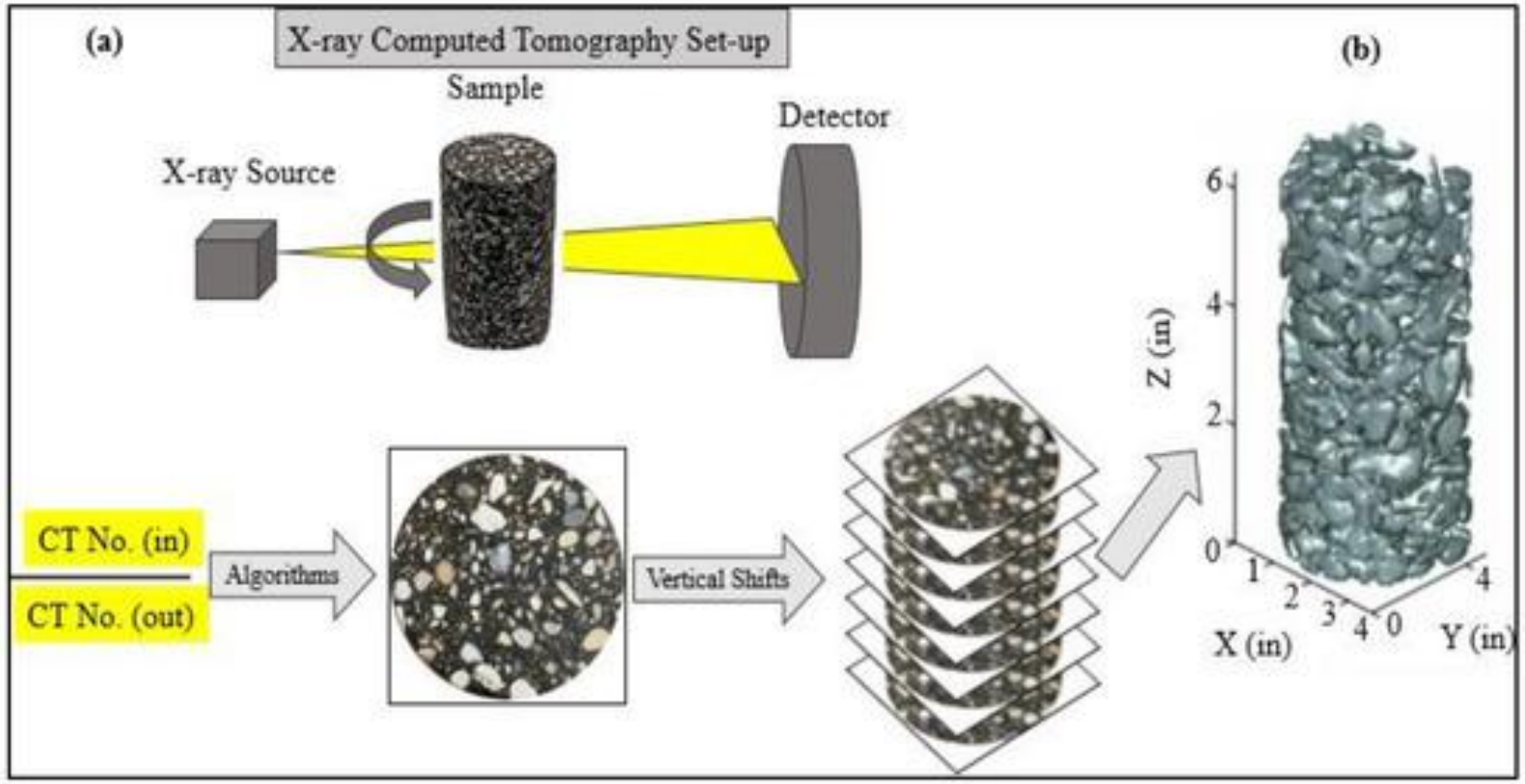
Fugaku

Use advanced compression and data management approaches

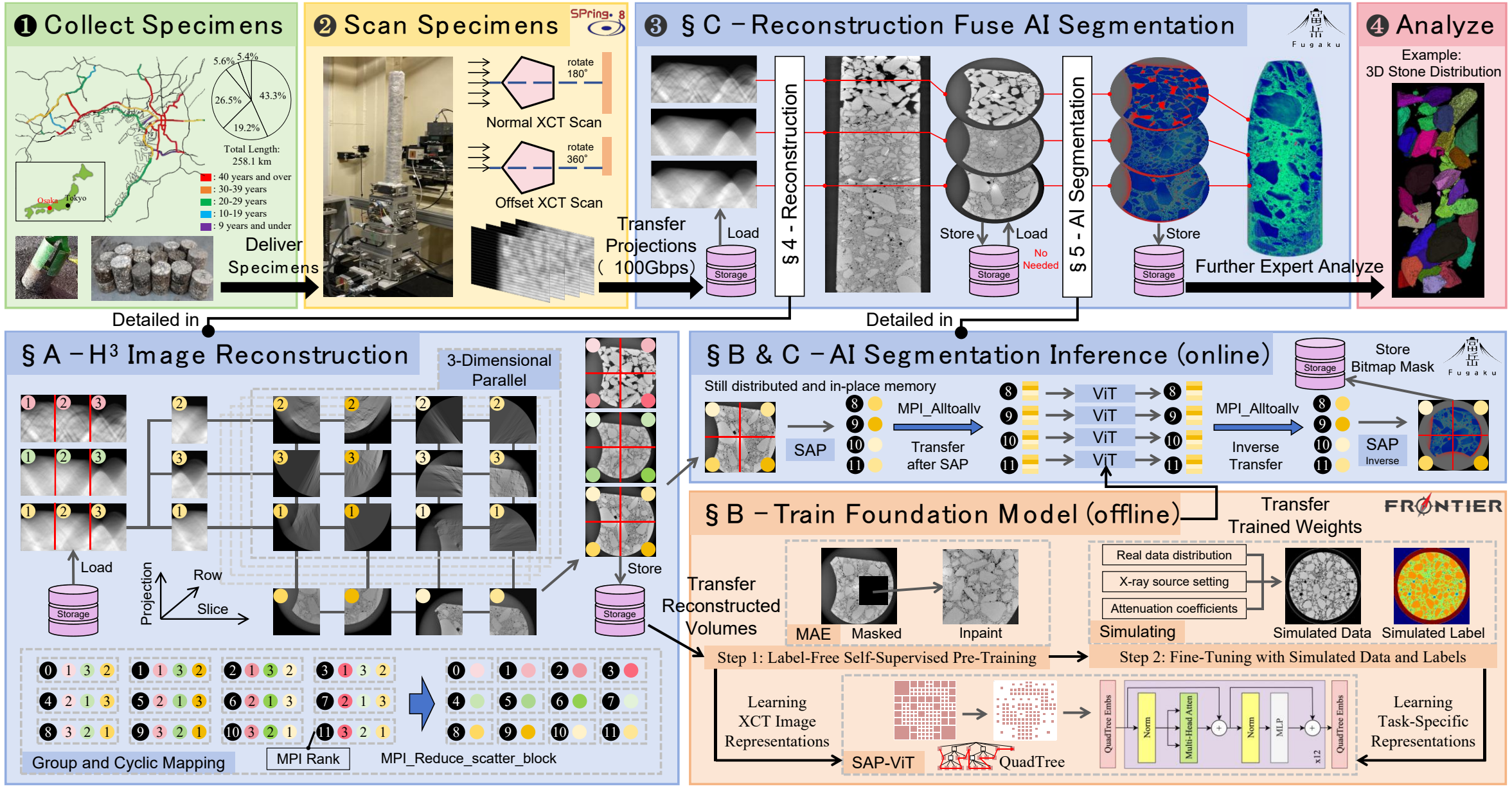
* *~2,000x amount of data used to train ChatGPT: Petabytes*

High Resolution 3D Image Reconstruction

- High-performance high-resolution X-ray CT image reconstruction



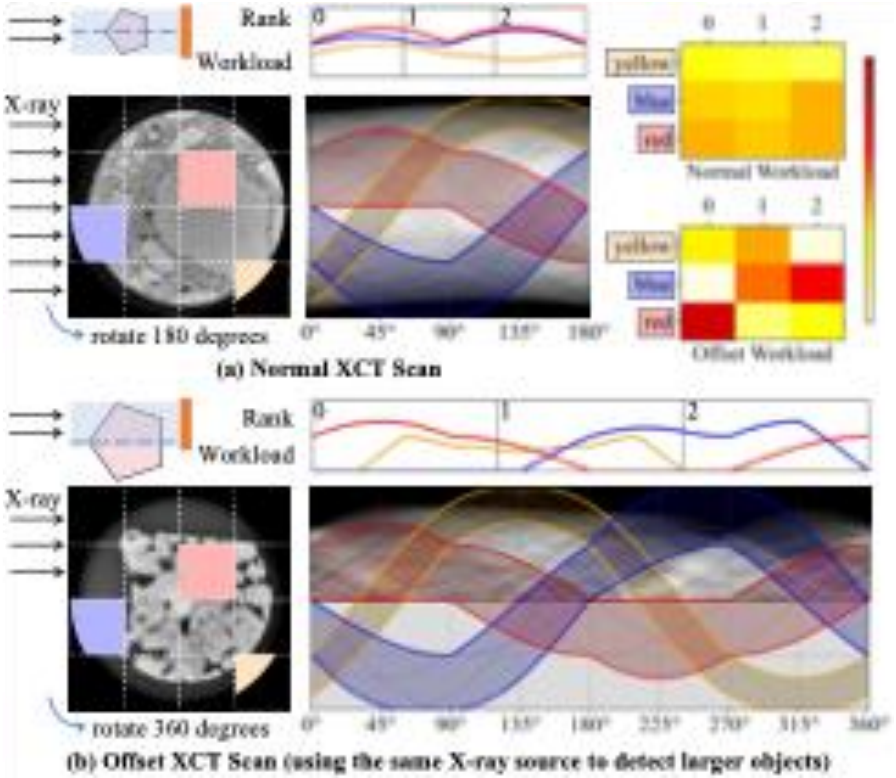
Paradigm Shift in Infrastructure Inspection Technology



H³: High-throughput, High-perf., High-resolution CT

H³

End-to-end pipeline to reconstruct 10s of 16K resolution 3D images in one go (full-system scale)



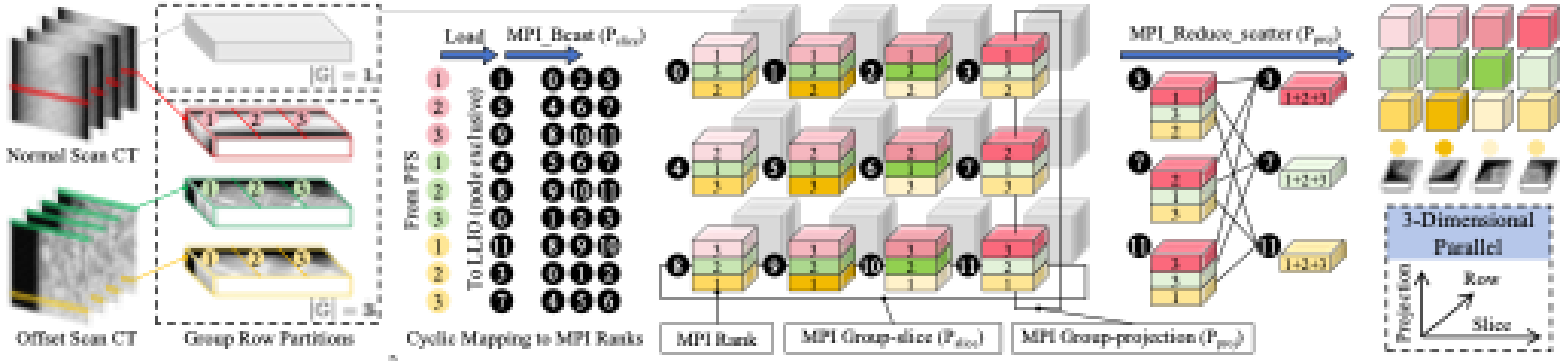
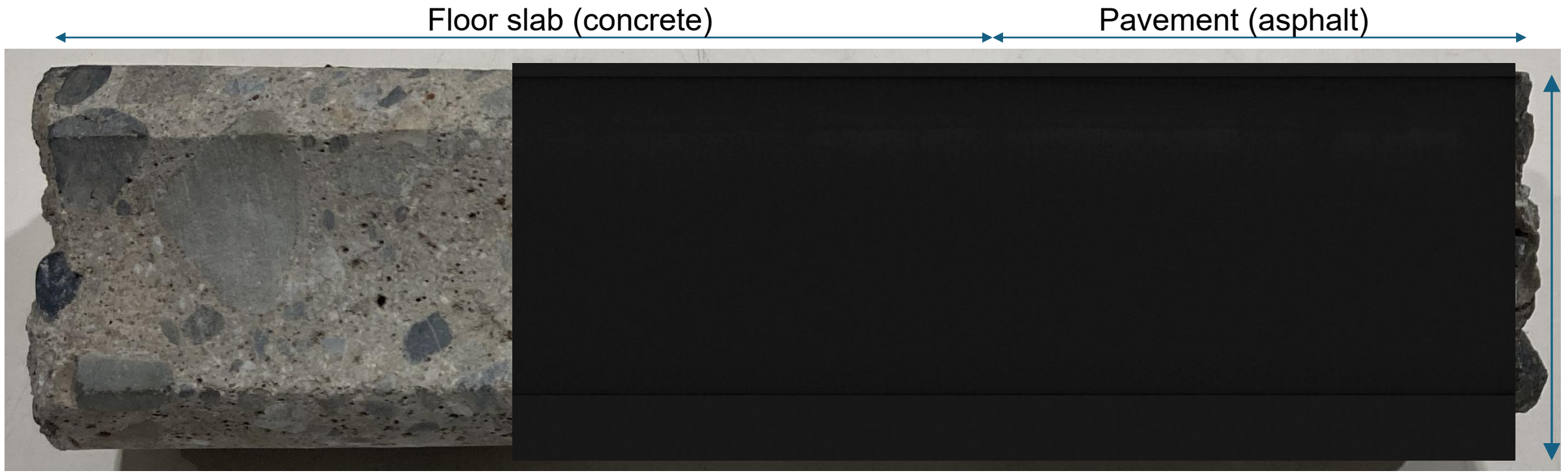


Fig. 5: Proposed H³ imaging algorithm in ROVAI: Group row partitions and cyclic mapping to MPI ranks.

End-to-end Image Reconstruction

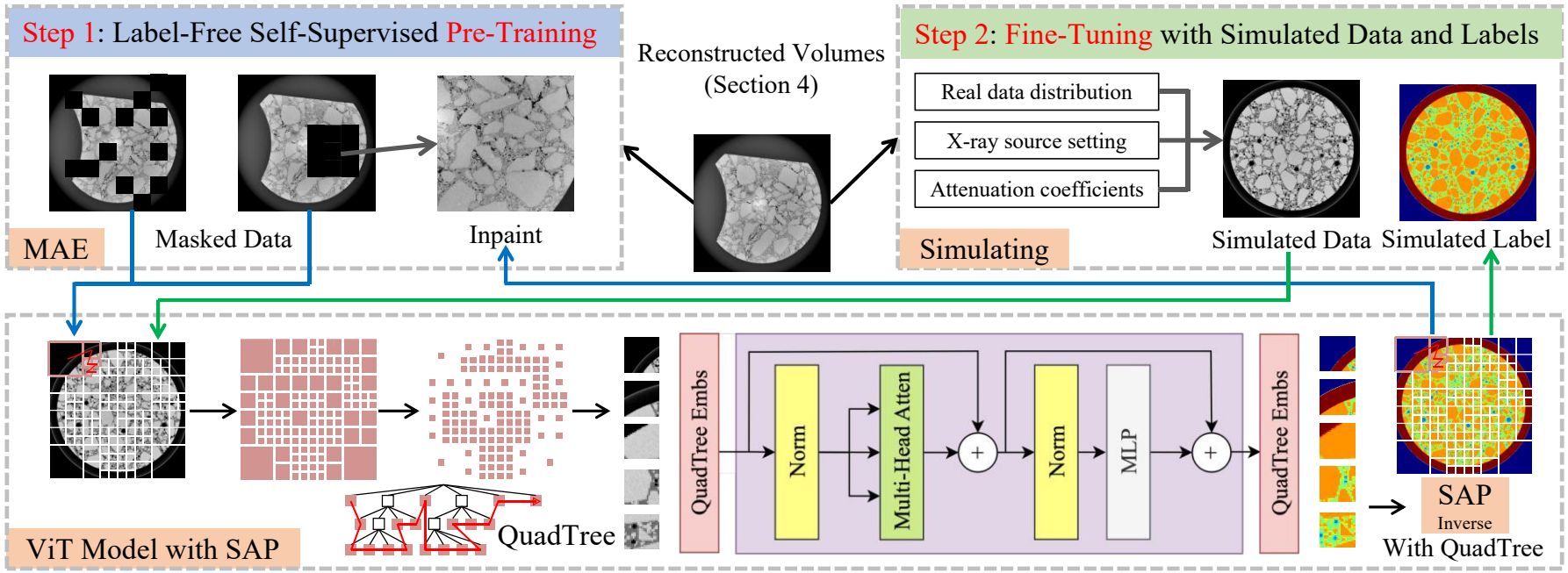


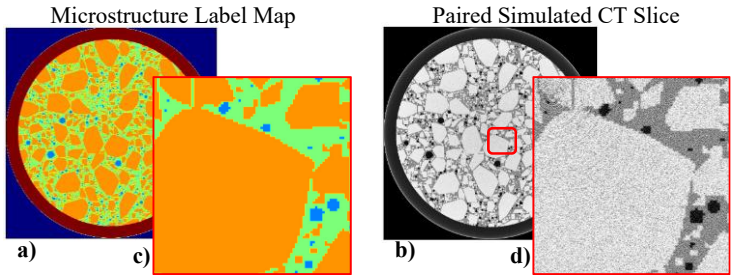
Cores were collected from road surfaces used for 20 to 30 years on the Hanshin Expressway.

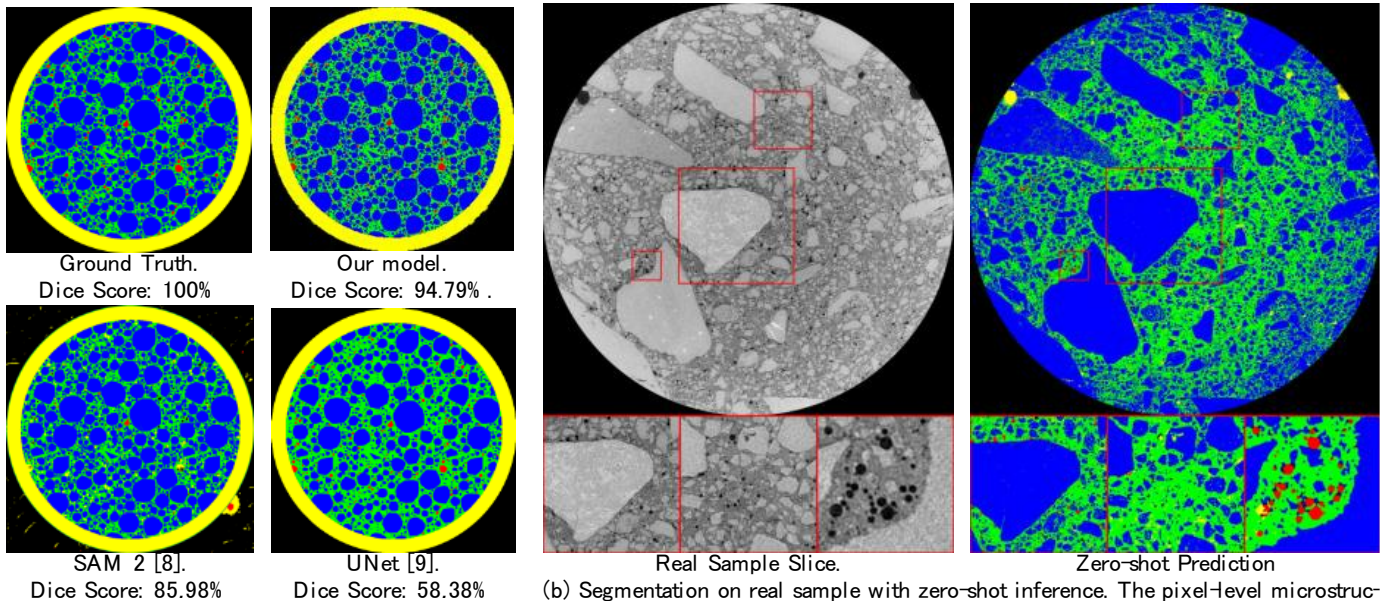
High-resolution $4{,}192^3$ Asphalt Core generated on Fugaku in 12 seconds (12,288 nodes: ~7% Fugaku)

* Courtesy of Kentaro Uesugi (RIKEN Spring-8)

Training a Foundation Model







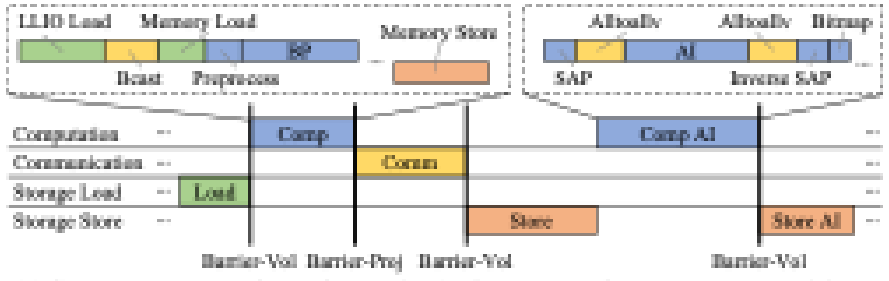
(b) Segmentation on real sample with zero-shot inference. The pixel-level microstruc-

Several Challenges

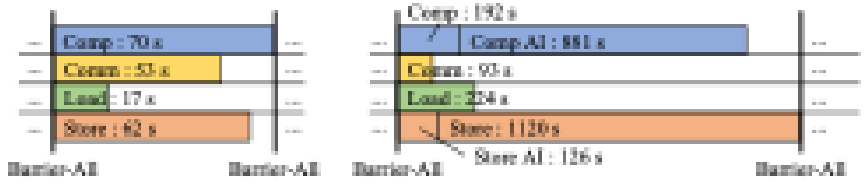
Sequence length, tokenization, shifting bottleneck, and **NO LABELED DATA**

Results

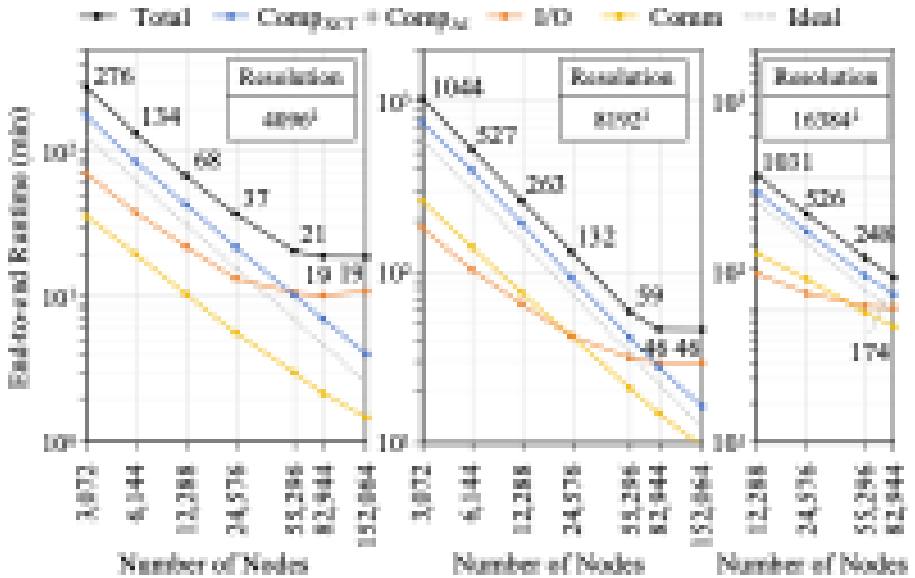
Performance



(a) Step-by-step workflow of the ROVAI framework for group row partitions.

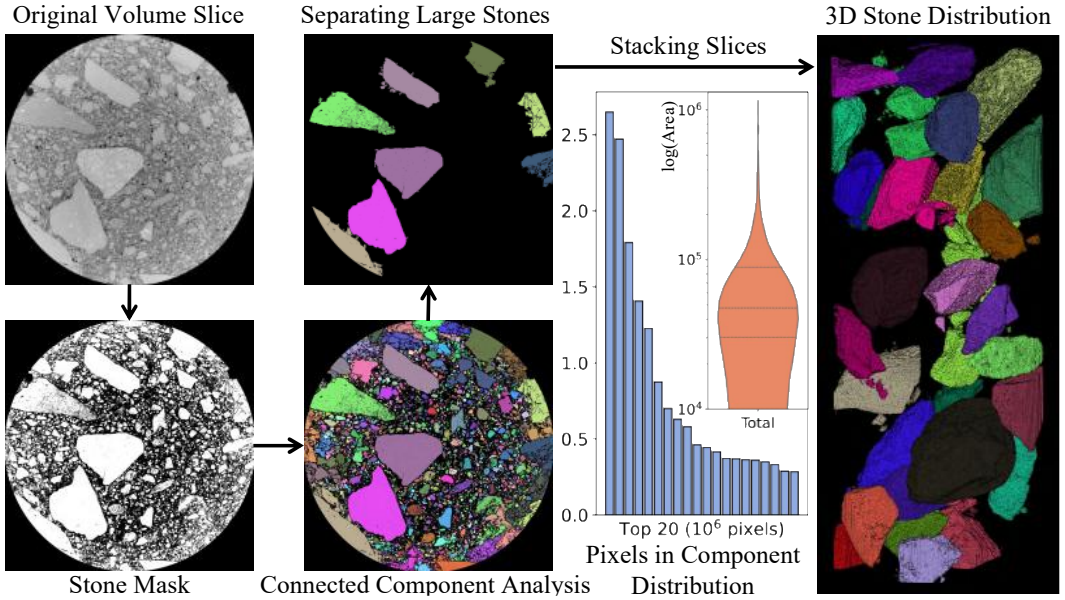


(b) Pipeline for H³ imaging. (c) Pipeline for entire ROVAI framework.



(a) Strong scaling evaluated by full Fugaku; ROVAI reconstructing 46 specimens into various resolution.

Analytics



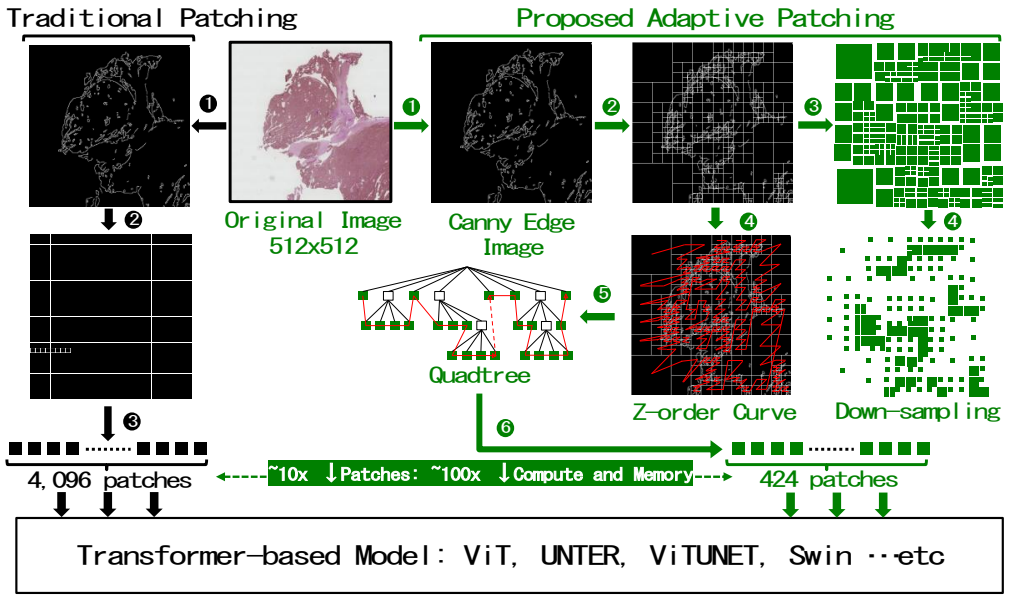
Road Inspection Analysis Items	T	XCT	XCT+AI
Detect surface degradation	✓	✓	✓
Visualize sub-surface conditions		✓	✓
Measure road shear strain angle		✓	✓
Measure density of concrete and aggregate		✓	✓
Measure the distribution of voids			✓
Measure the state of material around voids			✓
Measure volume ratio of aggregate, stone, etc.			✓

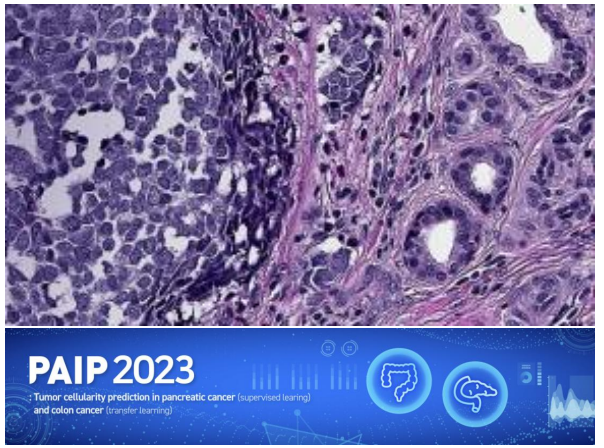
Datset	Model	Patch Size	GPU (hours)	Epochs	Dice (%)
780 unique volumes w/ simulated masks (8,192→8,192→(50←120))	U-Net [9]	N/A	1,280	500	58.38
	Swin UNETR [10]	256^2	5,120	1,000	63.74
	SAM 2[8]	128^2	5,120	1,000	85.98
	Our Model	2^2	5,120	1,000	94.79

AI Can Provide Real Value to Science

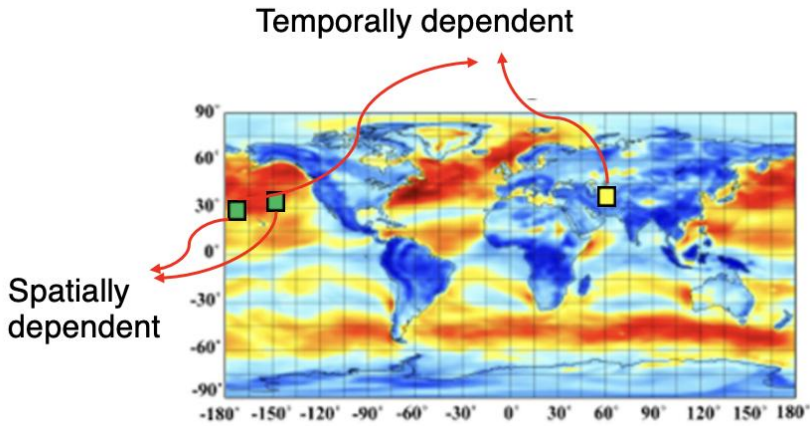
(High Performance Computing and Data management is a challenge)

microscopic pathology, Infrastructure Inspection, weather prediction

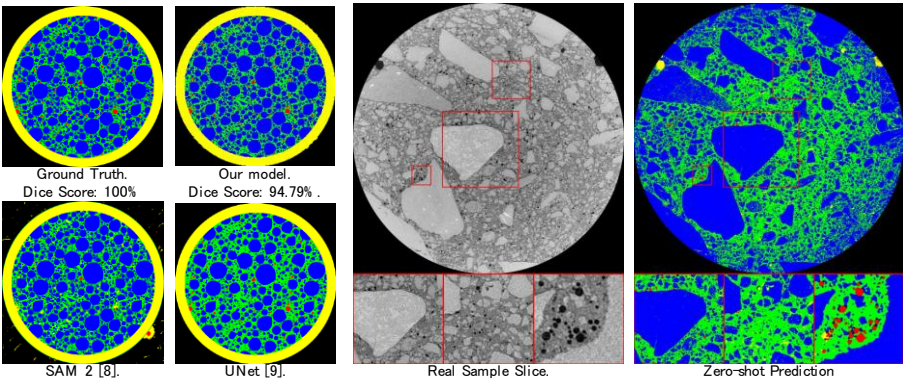


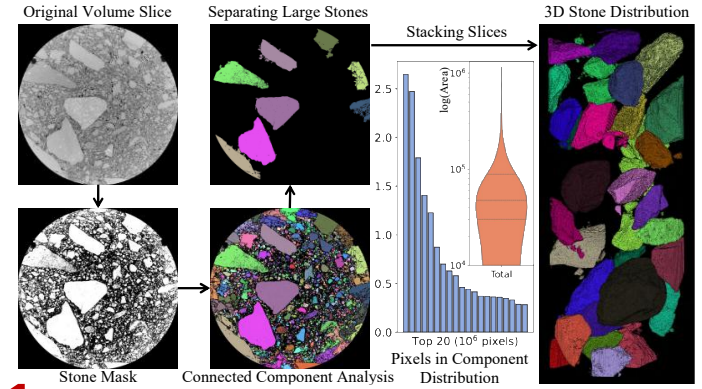


3

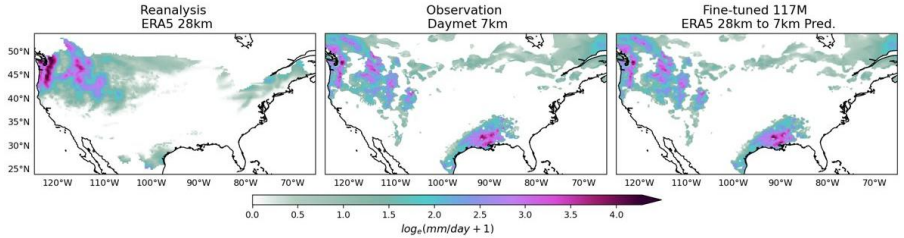


2





1



1 2 3

Thank you for listenting!