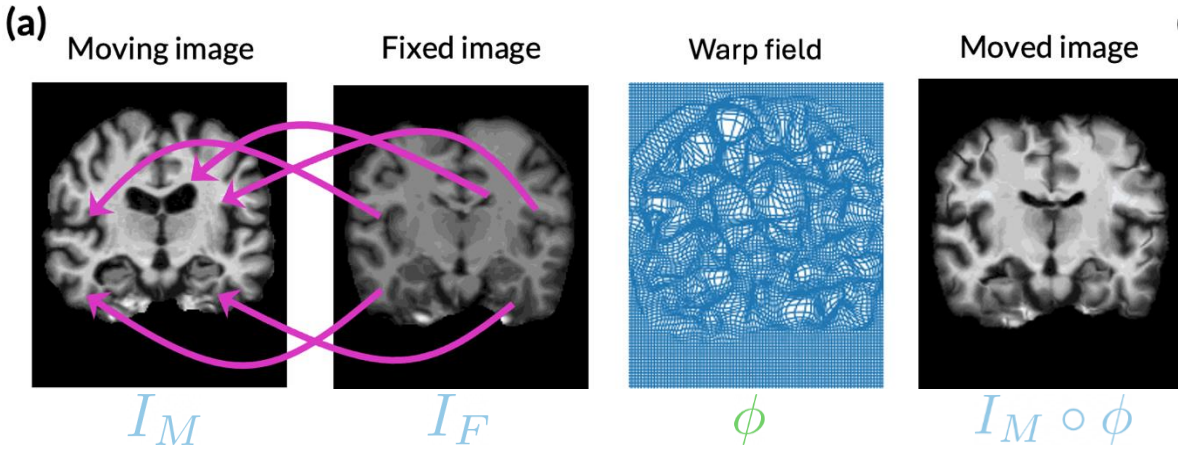
† Corresponding author
* These authors contributed equally

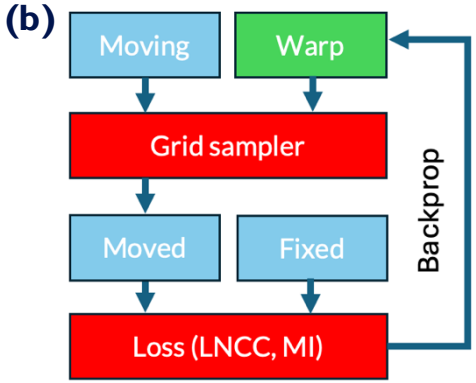
A Scalable Distributed Framework for Multimodal GigaVoxel Image Registration

Rohit Jena[†,*], Vedant Zope[*], Pratik Chaudhari, James Gee
{rjena, zope, pratikac}@seas.upenn.edu, gee@upenn.edu

ICLR 2026 Oral

Background

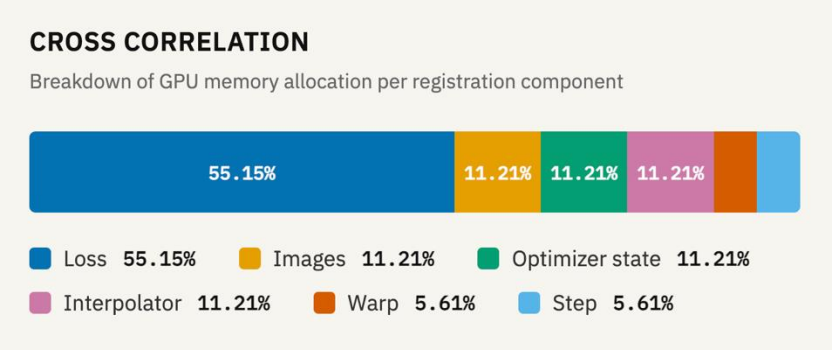


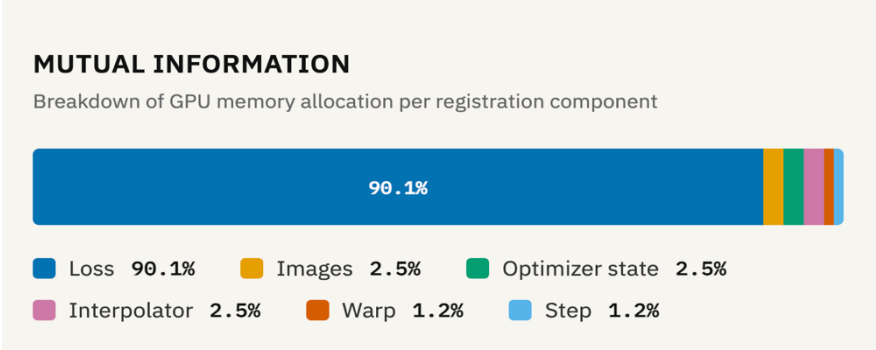


Process of establishing dense Corresponding Matching Between Fixed Image and Moving Image

$$\phi^* = \arg\min_{\phi \in \mathcal{G}} C(I_F, I_M \circ \phi) + \mathcal{R}(\phi)$$

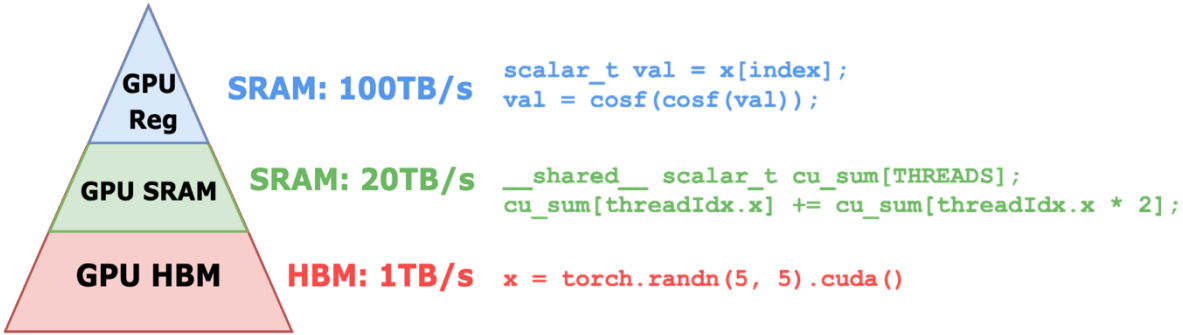
Memory Bottlenecks in registration





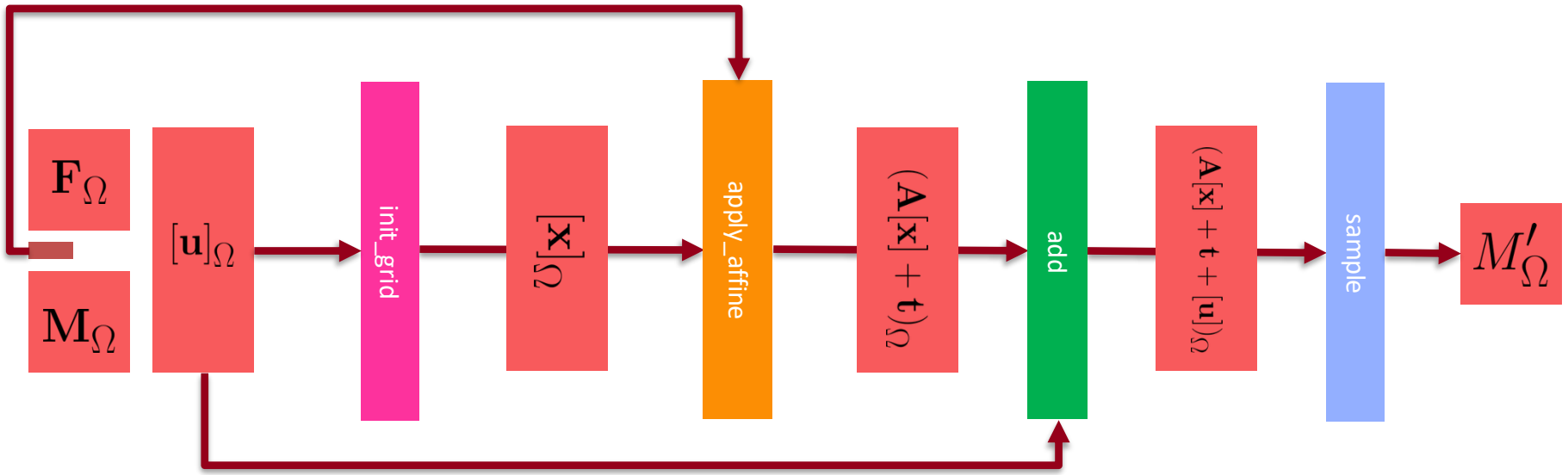
Intermediate tensors dominate runtime and memory

- Mutual information loss
- Cross-correlation loss
- Grid interpolation
- Warp composition



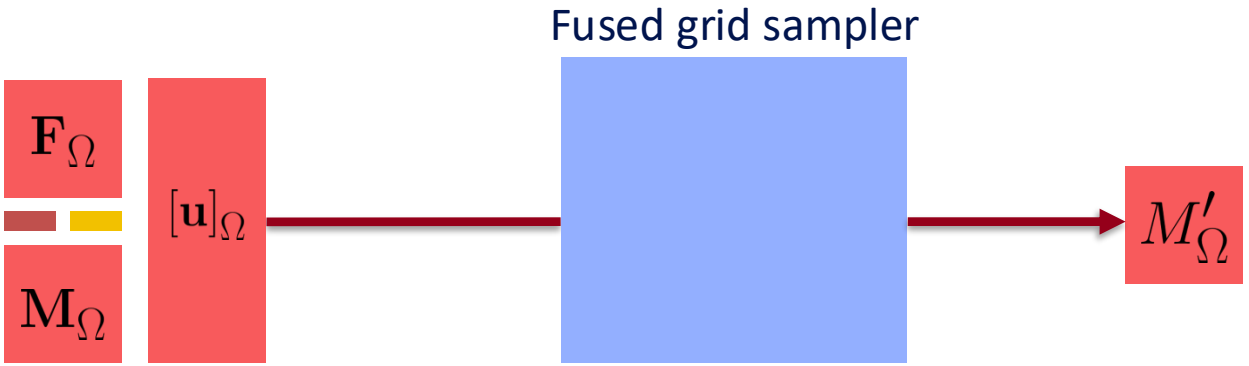
Grid Sampler/Interpolator

Composite Grid Sampler



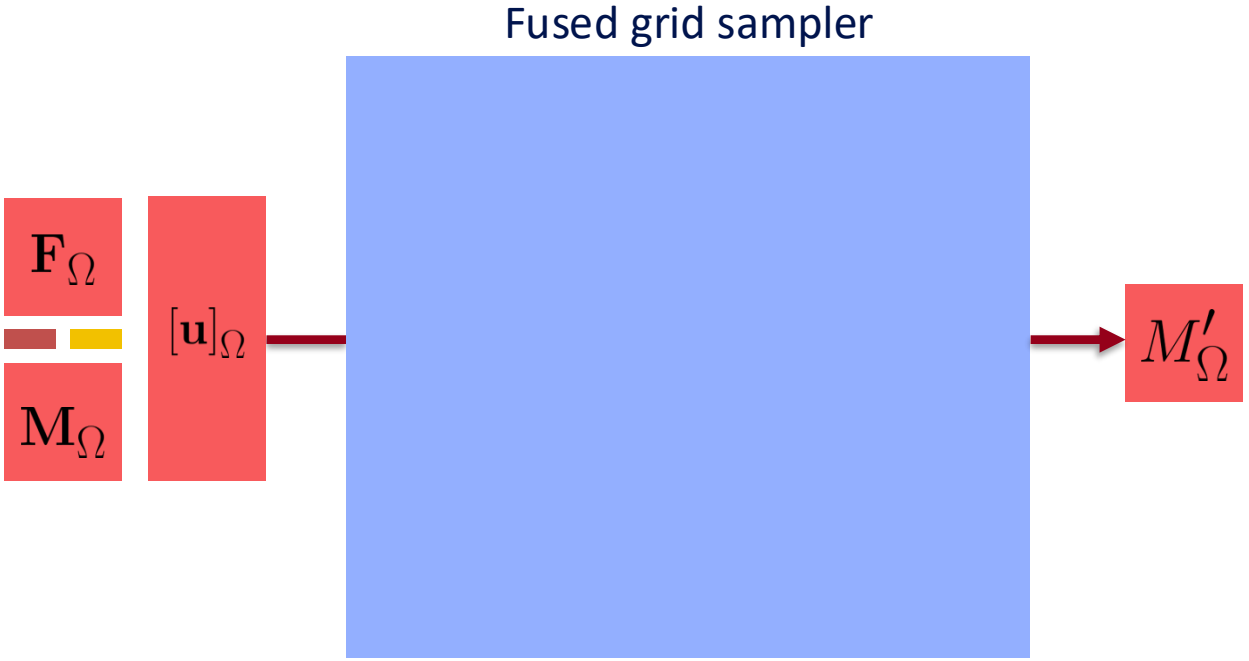
$$\texttt{grid\_sampler}(I;\ A, t, [\mathbf{u}])(x) = \texttt{sample}(I,\ \texttt{apply\_affine}(A,\ t,\ \texttt{init\_grid}(x)) + u)$$

Composite Implicit Grid Sampler



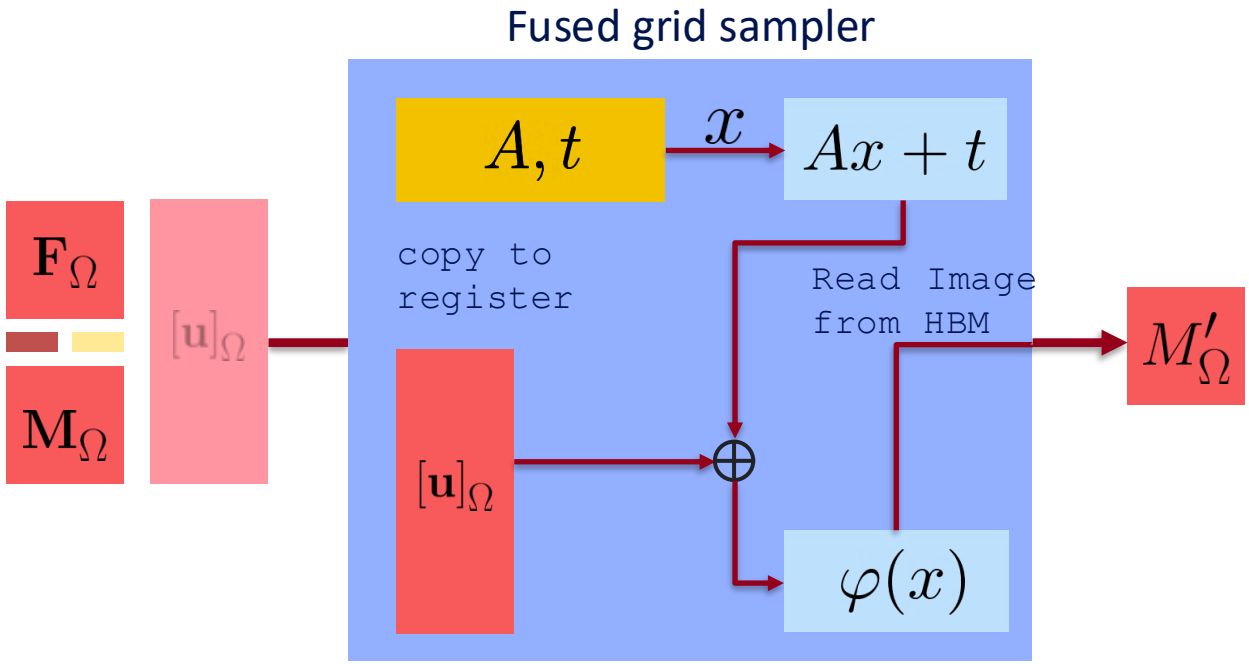
$$\texttt{fused\_grid\_sampler}(I; A, t, [\mathbf{u}], S, x_{\text{bounds}})(x) = I(Ax + t + Su(x))$$

Composite Implicit Grid Sampler



$$\texttt{fused\_grid\_sampler}(I; A, t, [\mathbf{u}], S, x_{\text{bounds}})(x) = I(Ax + t + Su(x))$$

Composite Implicit Grid Sampler



$$\texttt{fused\_grid\_sampler}(I; A, t, [\mathbf{u}], S, x_{\text{bounds}})(x) = I(Ax + t + Su(x))$$

Similarity Functions

Cross Correlation Loss

$\mathbf{F}_\Omega$ $\mathbf{M}_\Omega$

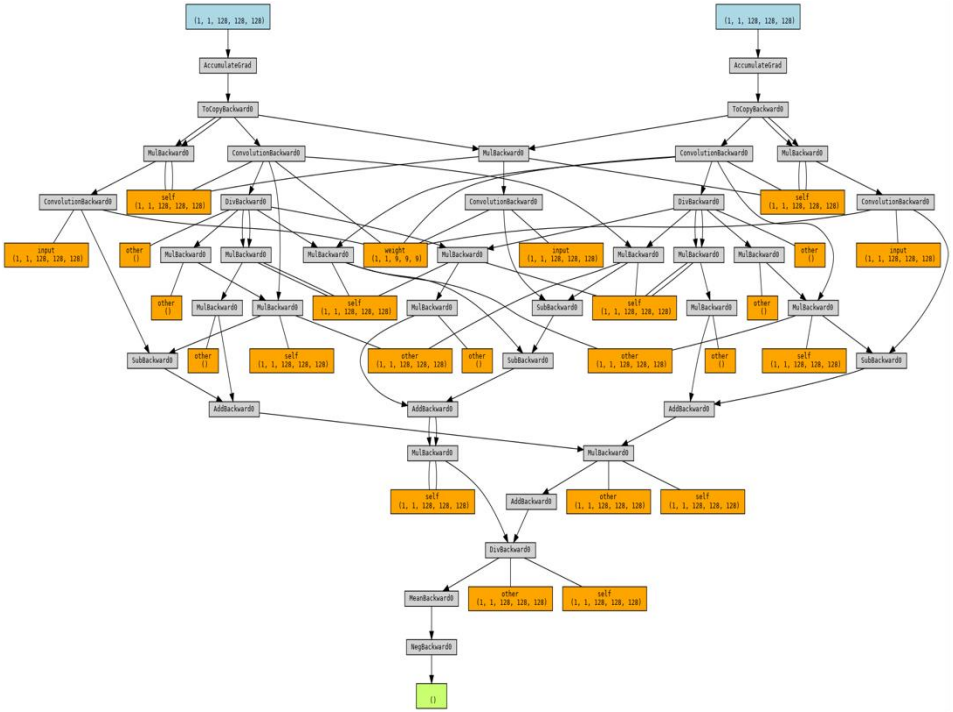
Vanilla LNCC

State Tensors + Intermediates

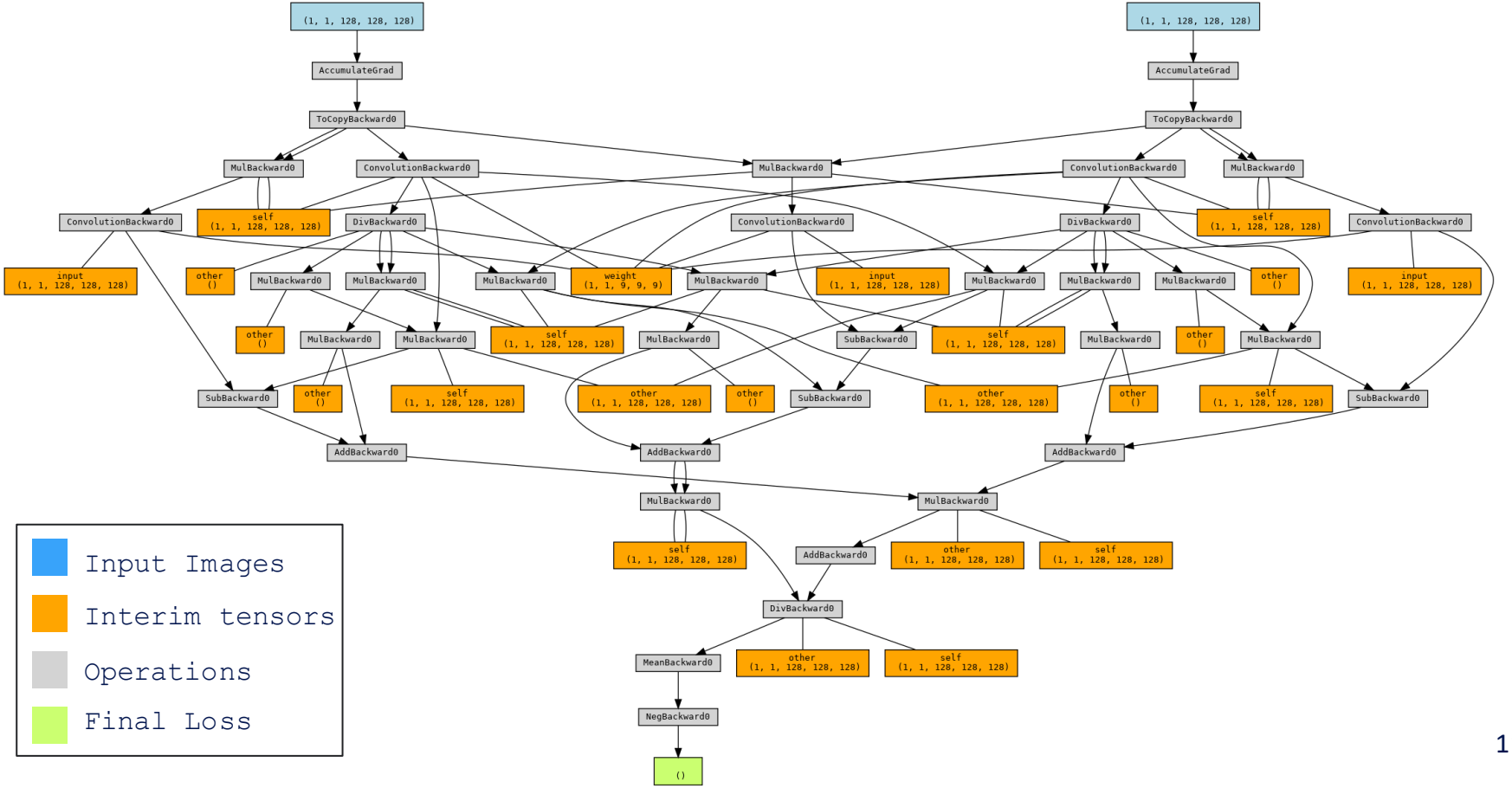
$\{F,\ M,\ F^2,\ M^2,\ FM,$
$\mu_F,\ \mu_M,\ \mu_F{}^2, \mu_M{}^2\ \ldots\}$

Stores:

- ~16N intermediates(forward)
- ~16N additional memory (backward)



Computational graph of the Vanilla LNCC



Cross Correlation Loss

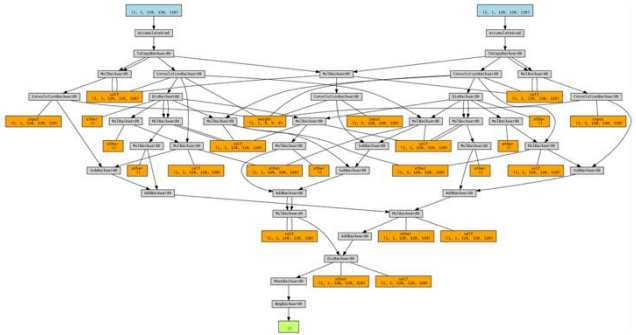
F_Ω M_Ω Vanilla LNCC

State Tensors + Intermediates

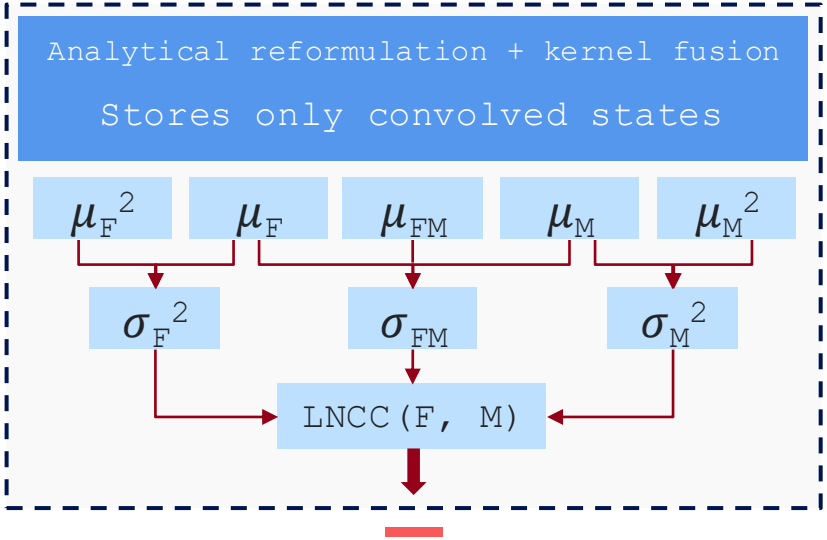
$\{F,\ M,\ F^2,\ M^2,\ FM,$
$\mu_F,\ \mu_M,\ \mu_F{}^2, \mu_M{}^2\ \ldots\}$

Stores:

- ~16N intermediates(forward)
- ~16N additional memory (backward)



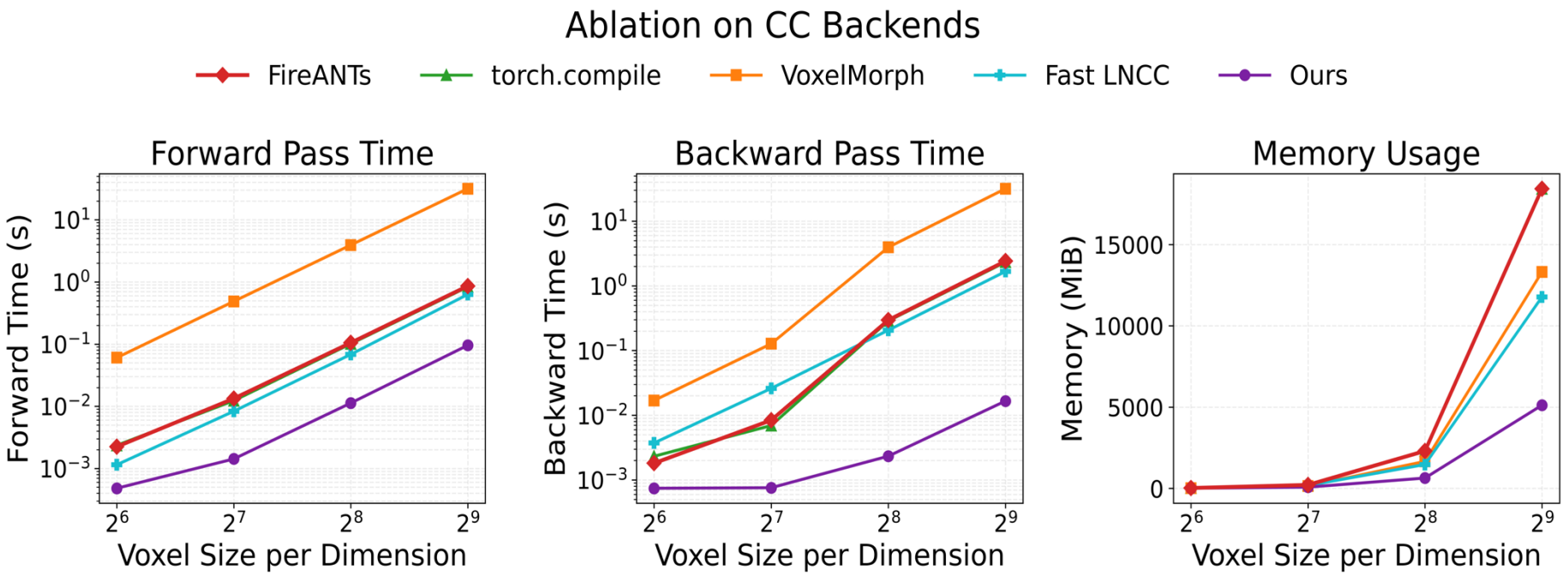
F_Ω M_Ω Fused LNCC



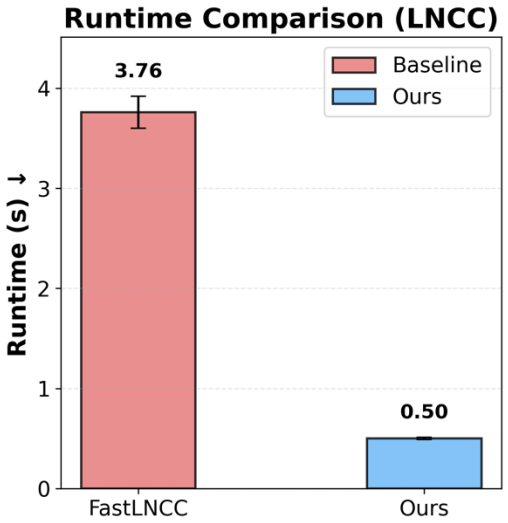
Stores:

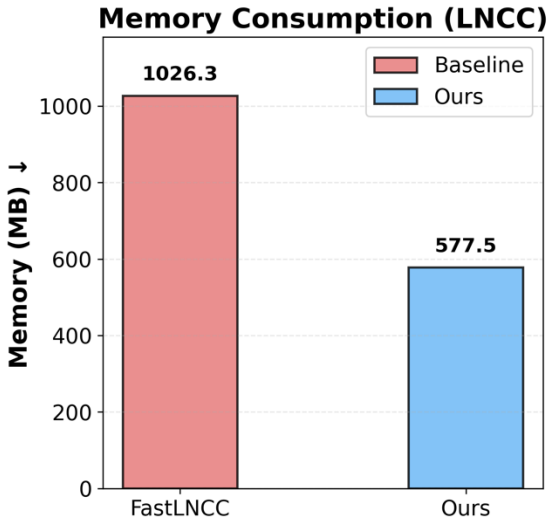
- ~5N intermediates (forward)
- No additional memory (backward)

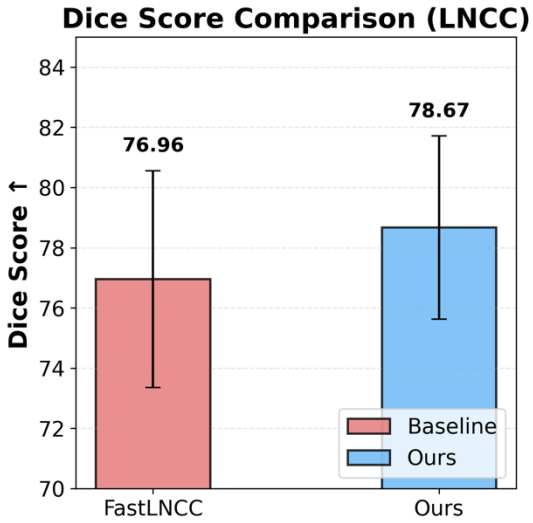
Fused Kernels Speedup (CC)



Iterative optimization becomes real-time

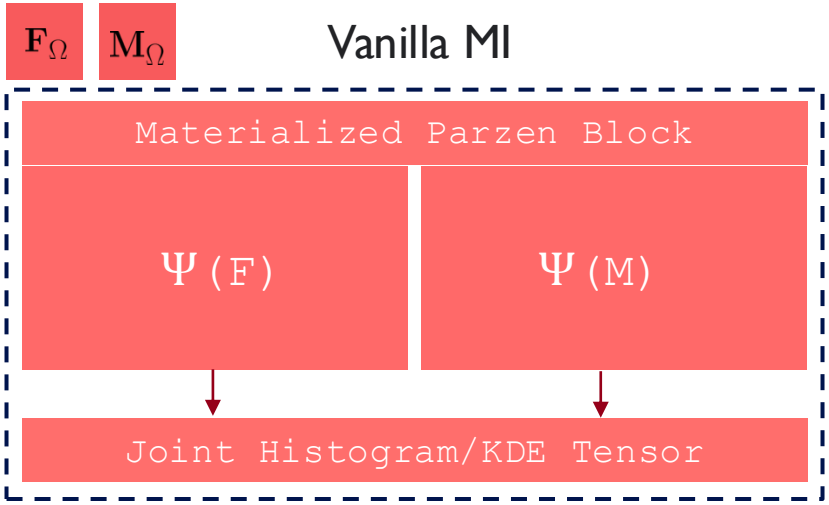




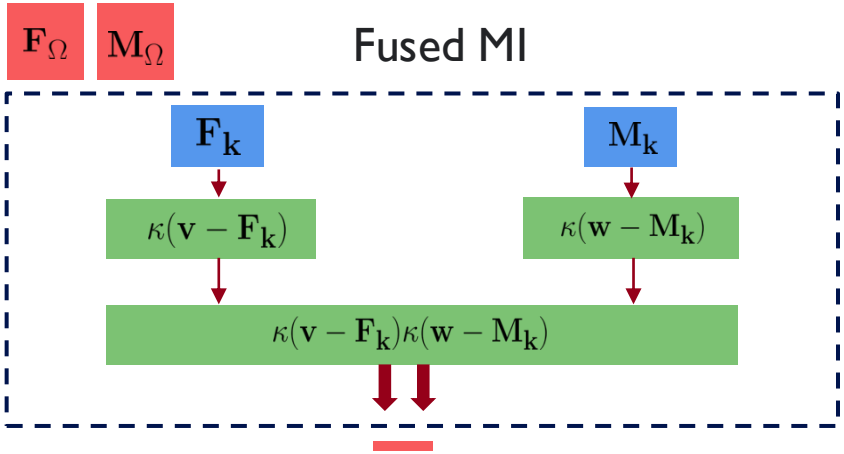


FireANTs + Fused Cross Correlation

Mutual Information Loss

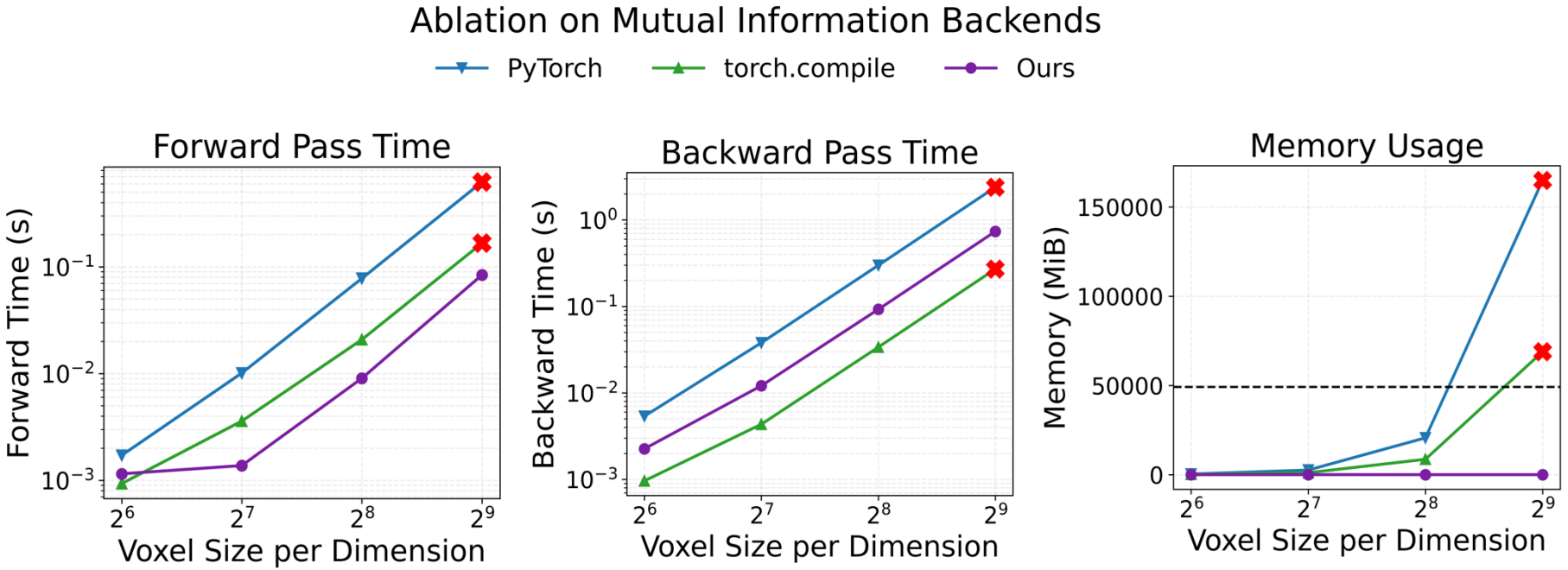


- Store full histogram in HBM
- Backprop requires storing all intermediates
- O(N) HBM overhead

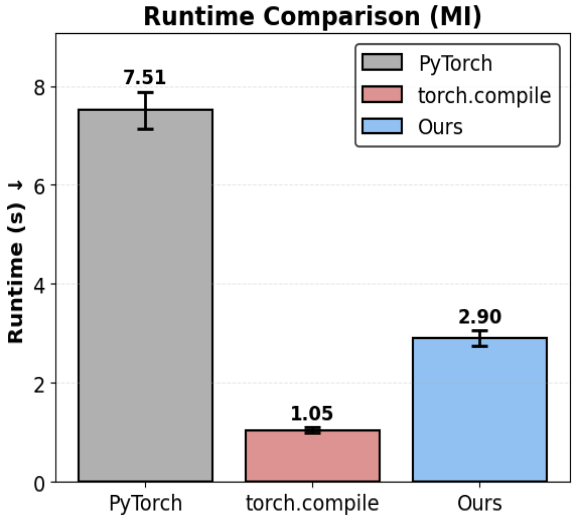


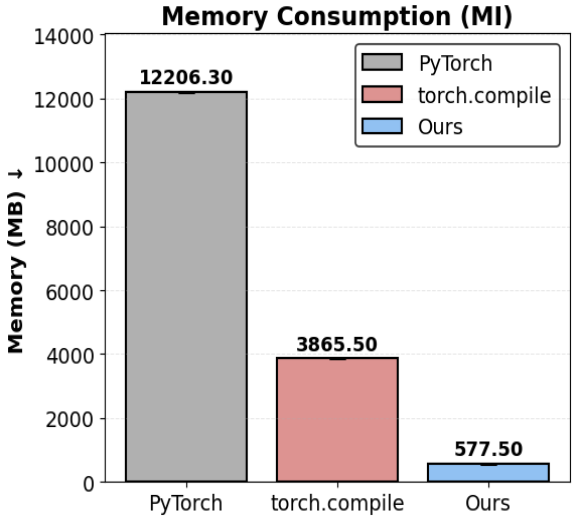
- Kernel contribution are computed per-thread and accumulated in shared memory
- Backward computed on-the-fly
- O(1) HBM overhead (holding B constant)

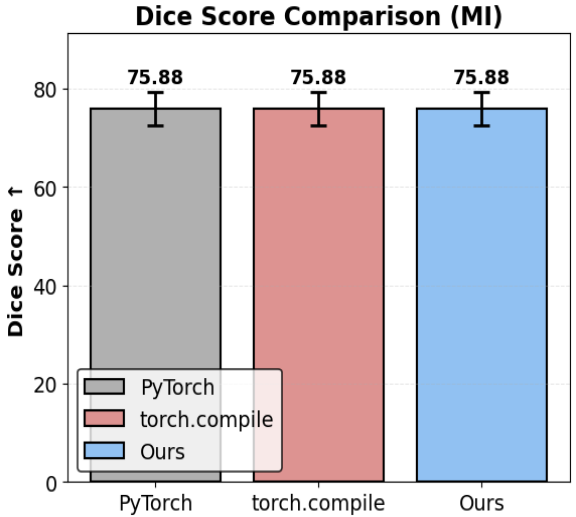
Fused Kernels Speedup (MI)



Iterative optimization becomes real-time

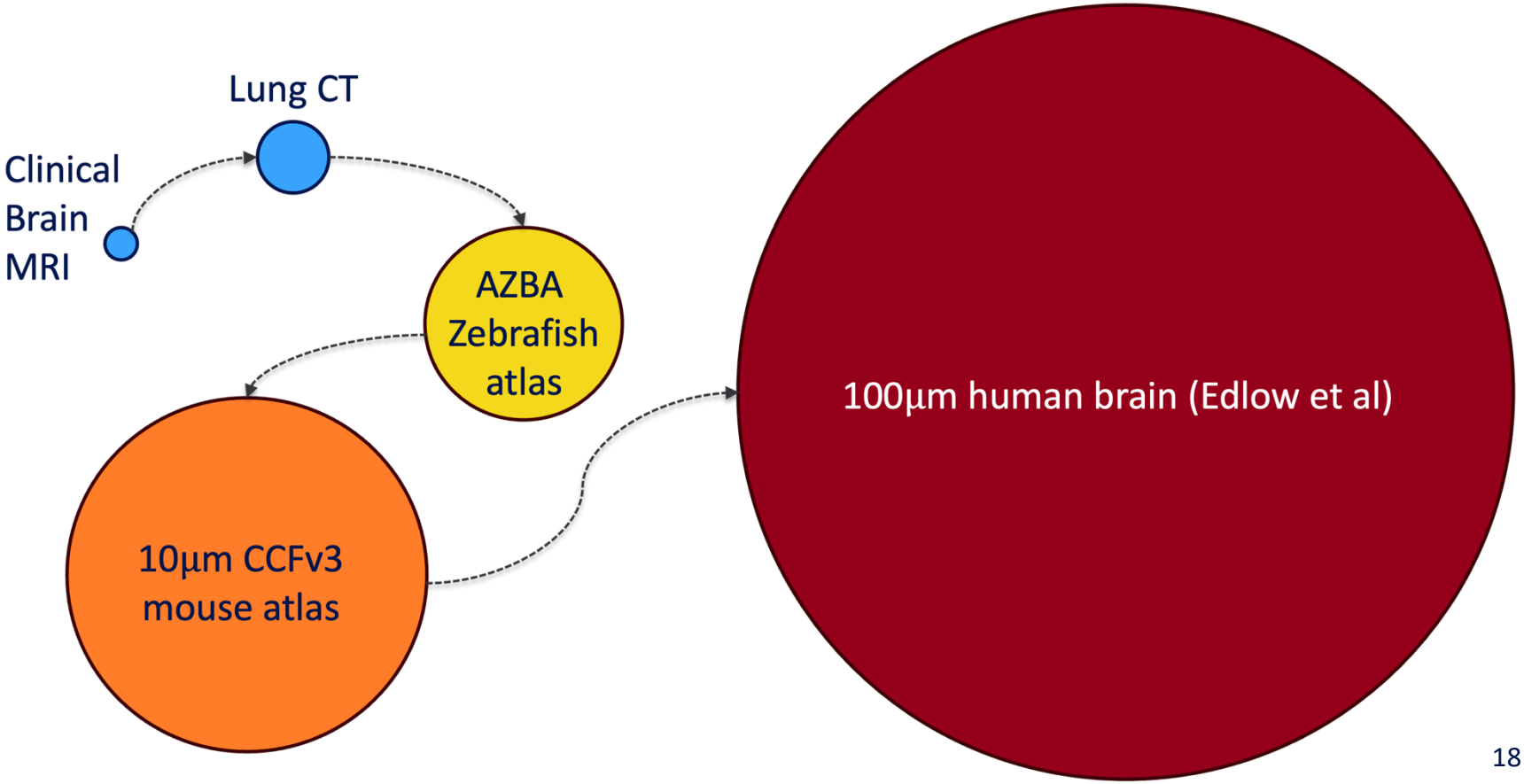






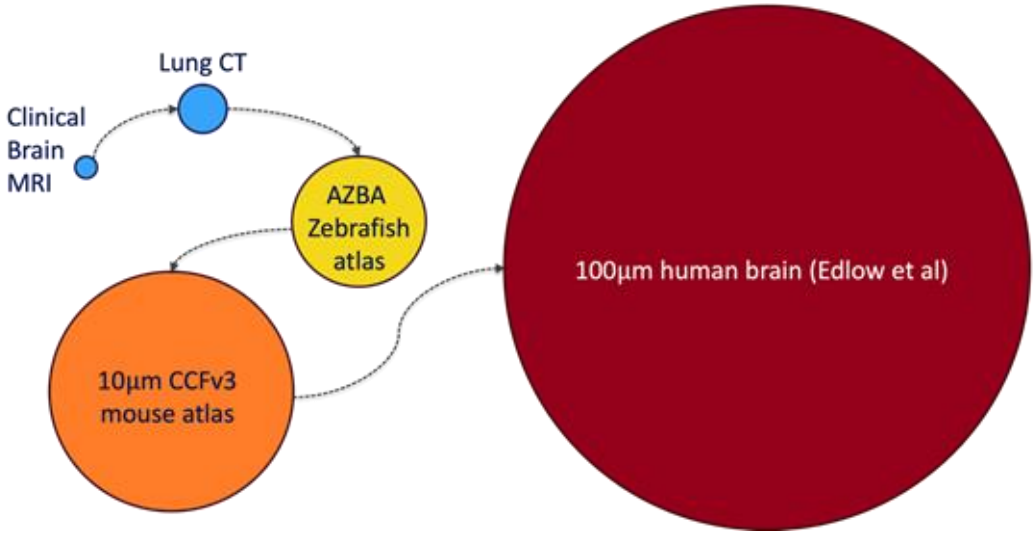
FireANTs + Fused Mutual Information

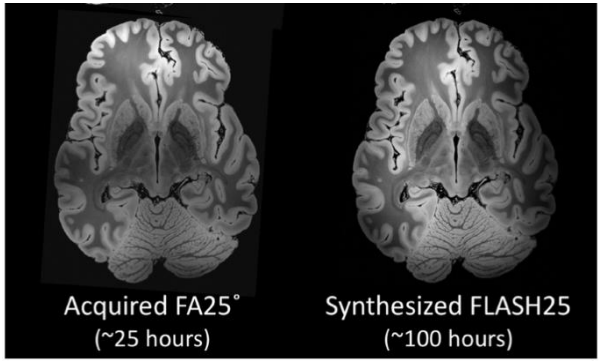
Imaging capabilities have scaled rapidly



Imaging capabilities have scaled rapidly

What happens when Image is too large?





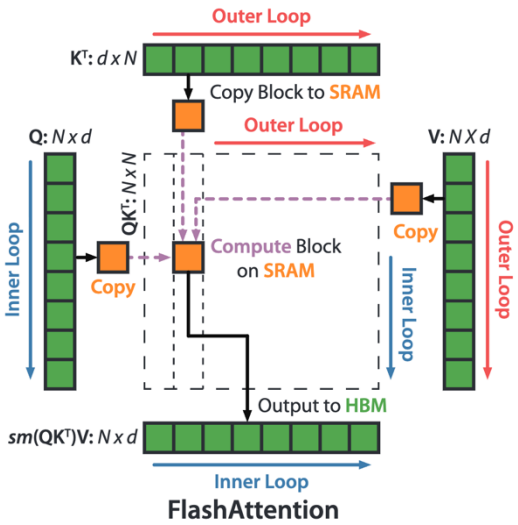
$[\mathbf{u}]_\Omega$ ≈ **44GB** VRAM

Edlow et. al. https://www.nature.com/articles/s41597-019-0254-8

Extending image registration to multiple GPUs

Background: Existing work

Two key ideas: Tiling, Kernel Fusion, Tensor Parallelism



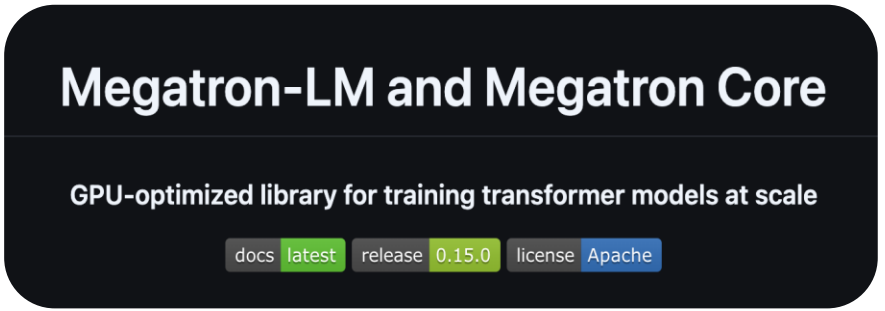
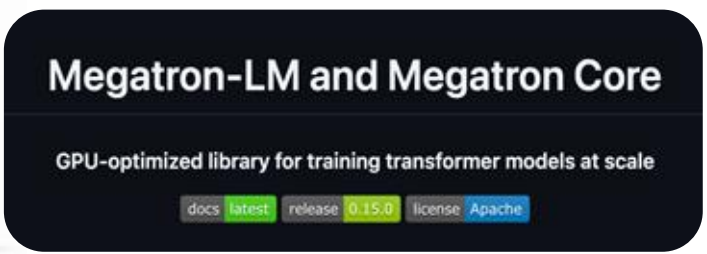


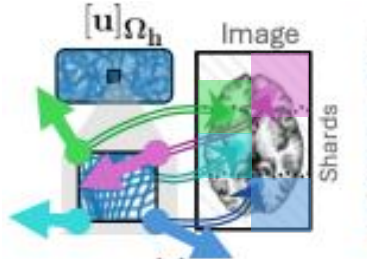
Image sources: flash-attn https://arxiv.org/pdf/2205.14135 , thunderkittens: https://github.com/HazyResearch/ThunderKittens

Background: Existing work

Two key ideas: Tiling, Tensor Parallelism

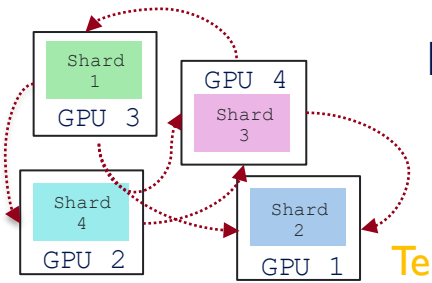


Why existing approaches don't scale?



Random access:

Tiling ✗



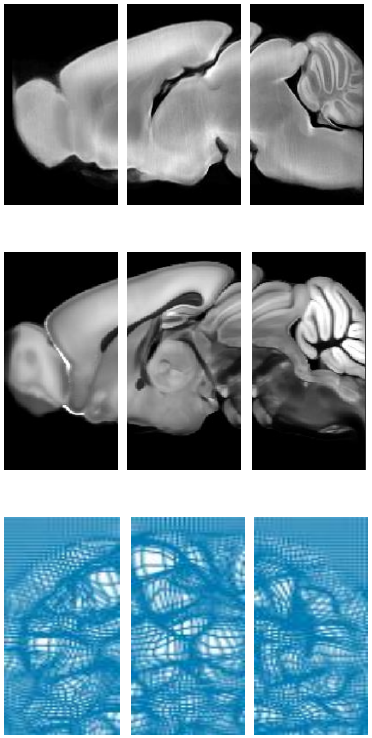
Random access + boundary sync:

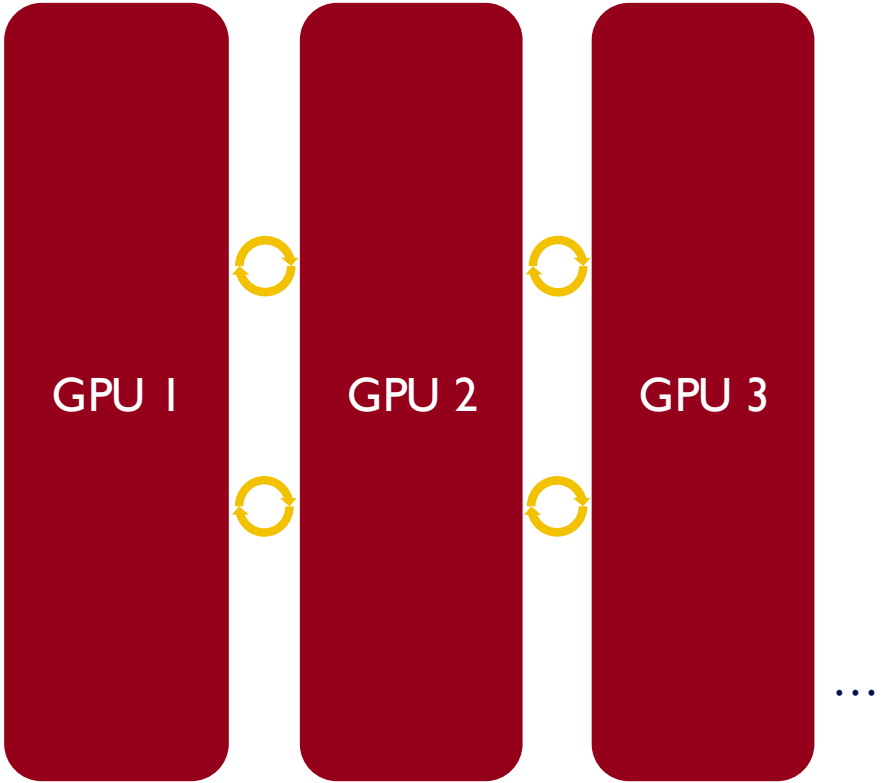
Tensor Parallelism ✗

Image sources: flash-attn https://arxiv.org/pdf/2205.14135 , thunderkittens: https://github.com/HazyResearch/ThunderKittens

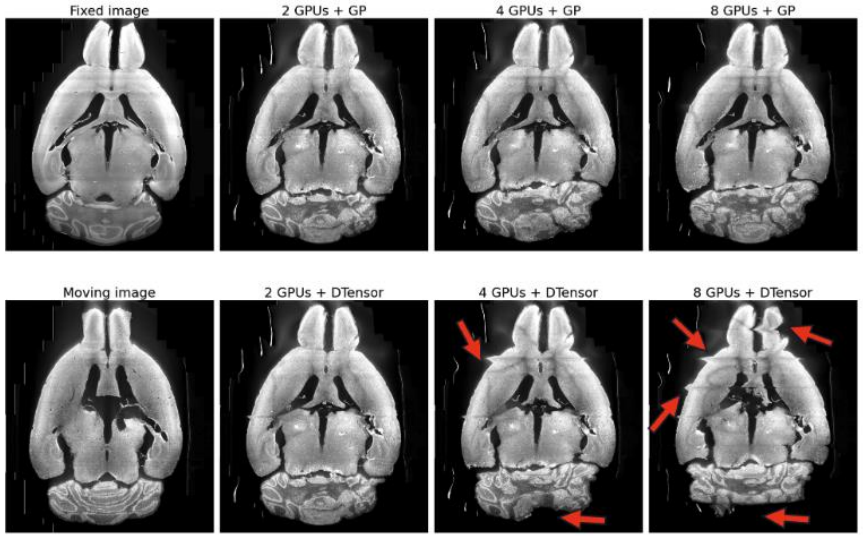
Grid Parallel

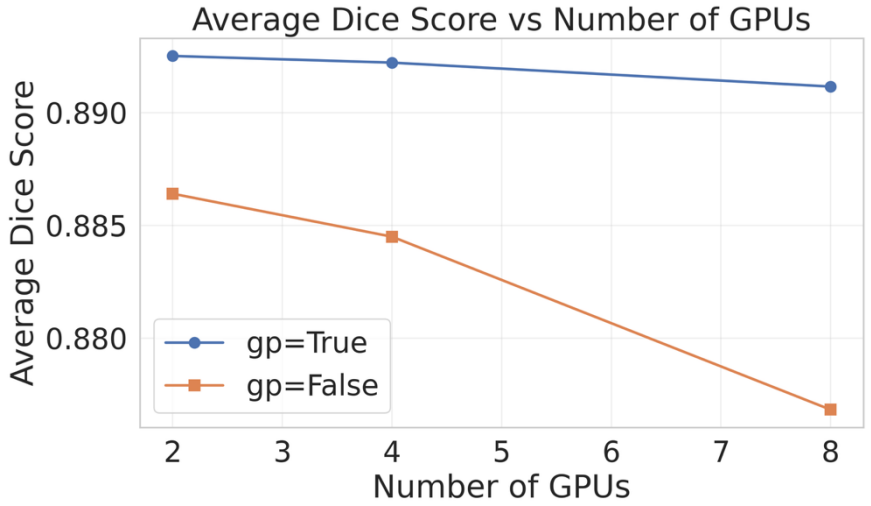
Grid Parallel





Why Grid Parallel?

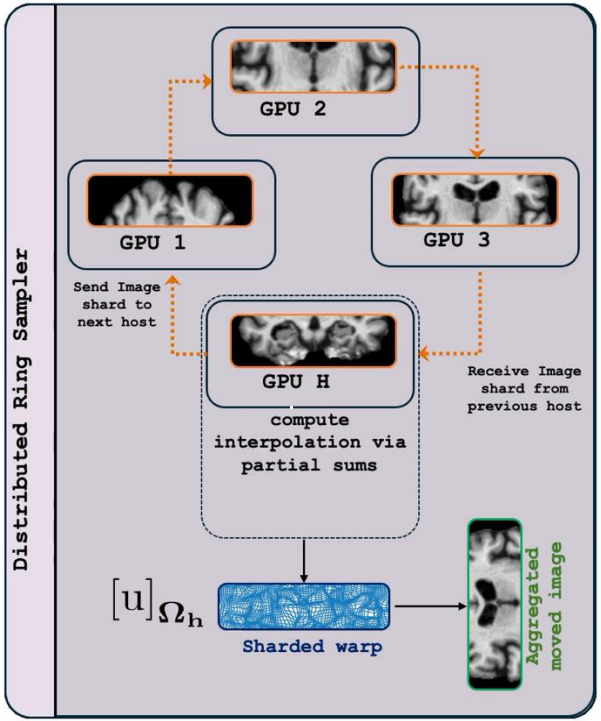




Distributed Ring Sampler

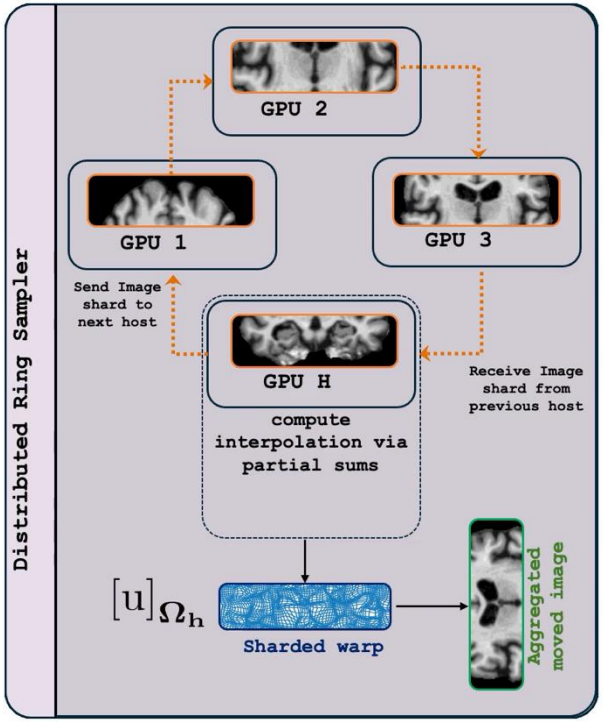
Distributed Ring Sampler

Ring Communication
(All-to-All via sequential passing)

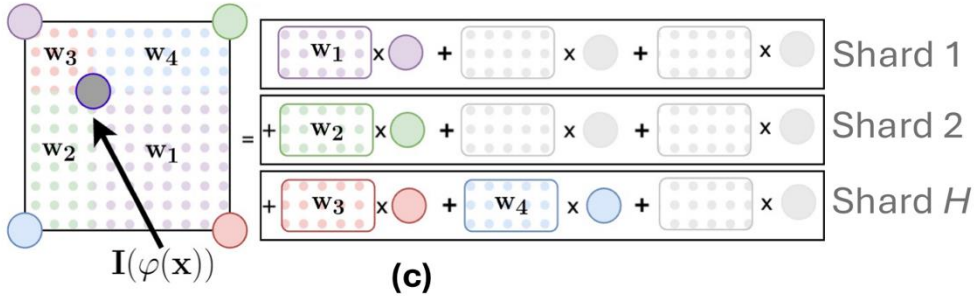


Distributed Ring Sampler

Ring Communication
(All-to-All via sequential passing)

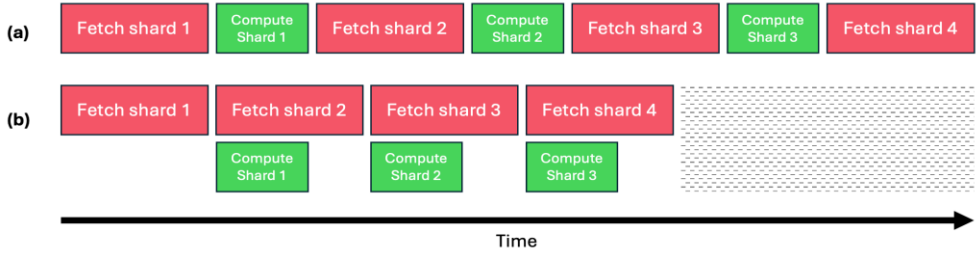


Interpolation requires cross-shard data dependencies

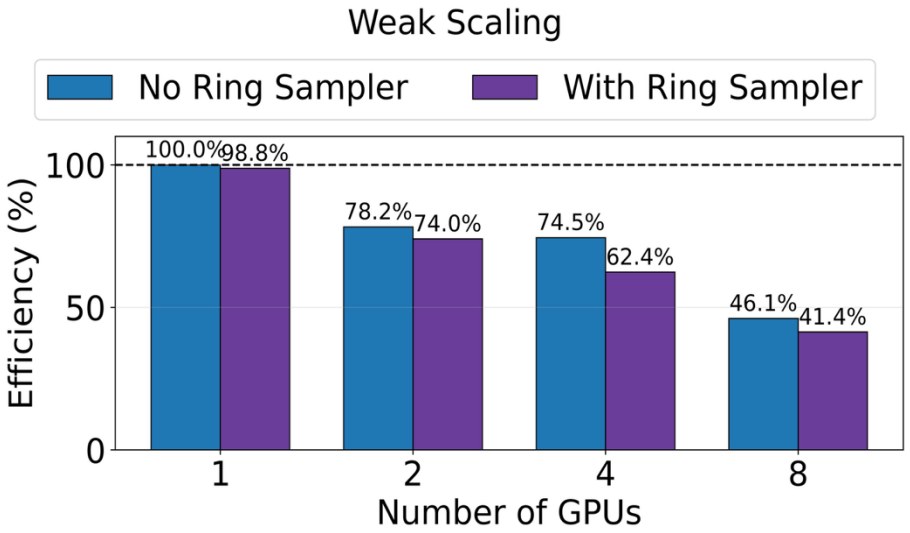


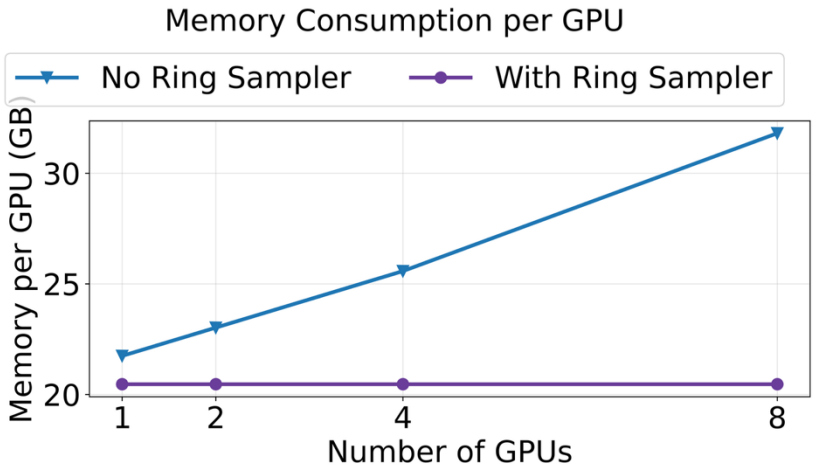
Partial sums → final interpolation

Each shard computes weighted contributions (w_i)

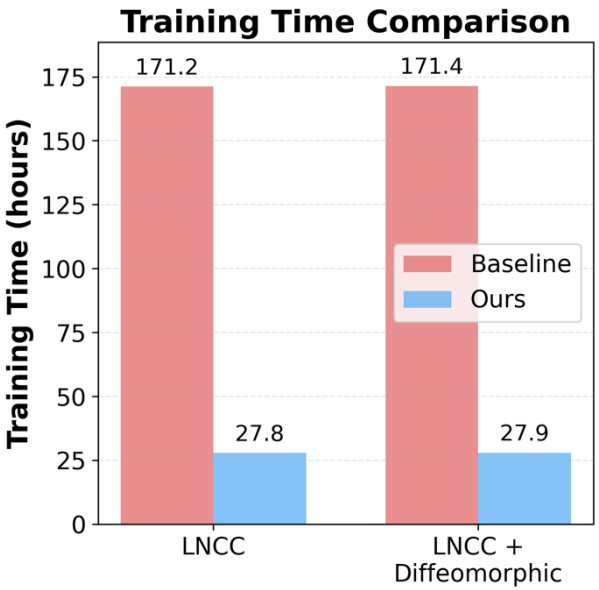


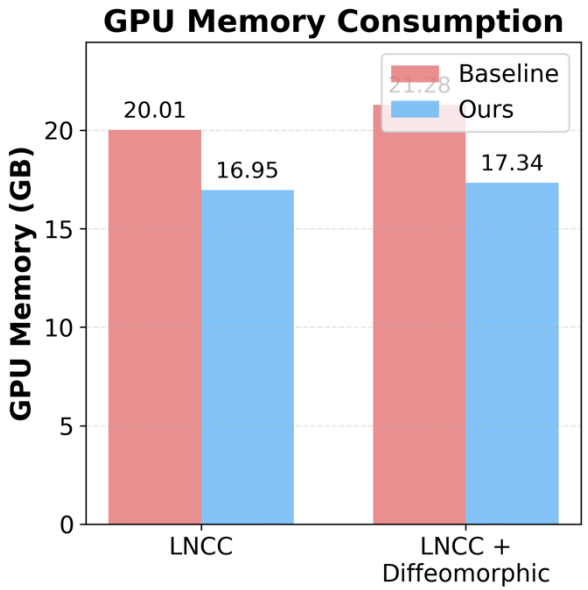
Scaling Results

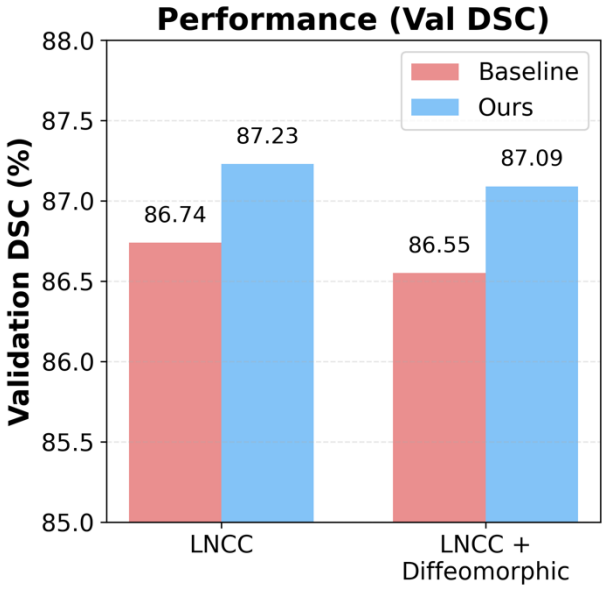




Deep Learning training speeds up by 6x







TransMorph

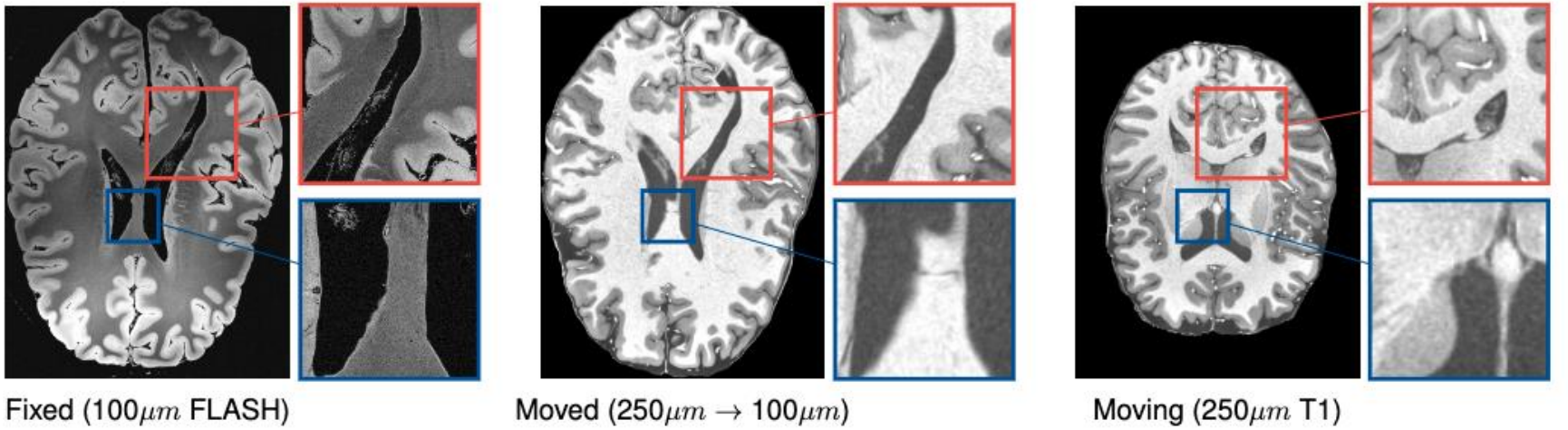
Multimodal Gigavoxel registration

Ex vivo (Edlow et al.) : **100 µm** FLASH imaging
1760 × 1760 × 1278 ≈ **11.8B** voxels

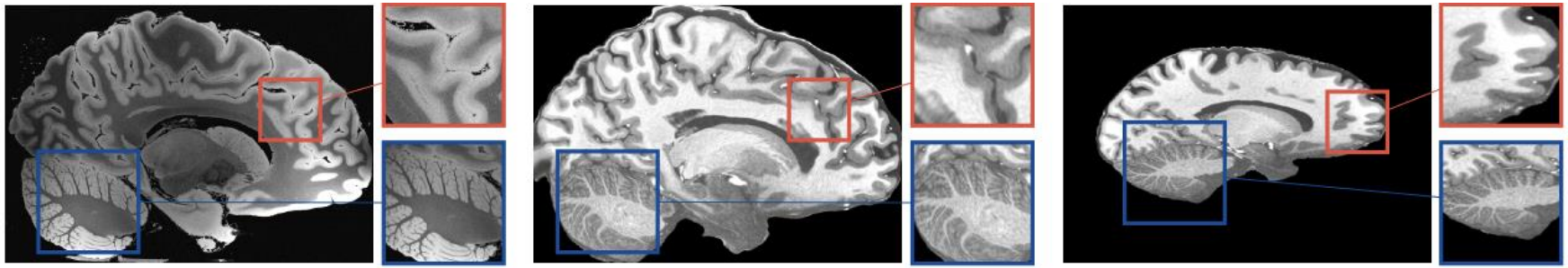
In vivo (Lusebrink et al.) : **250 µm** T1-weighted
800 × 800 × 800

Using **8** A6000 GPUs in ~**1 minute**

Qualitative Results



Fixed (100μm FLASH) Moved (250$\mu m \rightarrow$ 100μm) Moving (250μm T1)



Fixed (100μm FLASH) Moved (250$\mu m \rightarrow$ 100μm) Moving (250μm T1)

Thank you!

Q&A

Welcome to join me for a discussion during the poster session!

🗓️ April 25, 2026,

⏰ 2:15 PM – 4:45 PM (EDT)

📍 Poster Session 6