NEURIPS 2026 · SYDNEY, AUSTRALIA

Geometry Matters in Packed 3D Attention

Amil Khan · Anantajit Subrahmanya · B.S. Manjunath
Department of Electrical & Computer Engineering · University of California, Santa Barbara
amil@ucsb.edu

TAKE-HOME

Where a token is stored in memory should not change what attention sees.

We reorder the tokens of 3D cell images in memory, leaving the images untouched. A model that takes positions from memory order (sequence-index RoPE) loses 0.048 Dice on average, with a loss in all six seeds. PACK3D, our attention layer, takes them from 3D coordinates and changes by 0.0001.

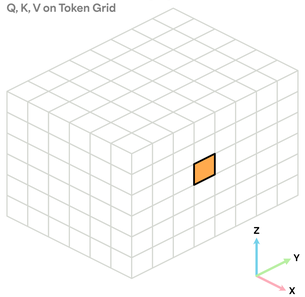
Scan for the paper
arXiv:2610.38694

THE SETUP

A 3D image becomes a sequence of tokens

We segment 15 structures in 3D images of live human stem cells. Each image is cut into tokens (patches of 1 × 4 × 4 voxels) and laid out in memory in an order chosen for speed: row by row, by blocks, or along a space-filling curve. Code should be free to change that order.

1. 3D Token Field
Q, K, V on Token Grid

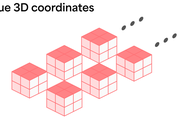


2. Pack for Memory (layout only)

Pack tokens for memory

Block-Morton / block-contiguous layout

Preserve true 3D coordinates

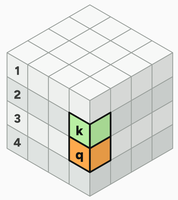


3D Grid Coordinates (z, y, x) travel with each token

THE PROBLEM

So every token has two addresses, and attention can learn the wrong one

Its place in the volume never changes. Its place in memory depends on the layout.

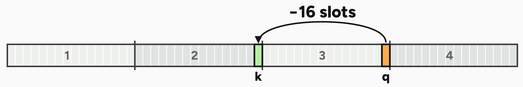


q a query token

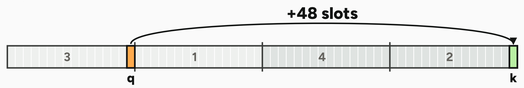
k one step above q

A 4 × 4 × 4 volume stored as four blocks of 16 tokens, one layer each

Training layout blocks 1, 2, 3, 4



Repacked same blocks, new order: 3, 1, 4, 2



	3D offset of k, (z, y, x) what PACK3D reads	Memory offset of k what sequence-index RoPE reads
Training layout	(1, 0, 0)	−16
Repacked	(1, 0, 0)	+48

Sequence-index RoPE, rotary position embeddings (RoPE) indexed by each token's place in memory, makes attention depend on memory offsets, so a model can tie part of what it learns to the layout. Because memory position is a consistent proxy for geometry on the training layout, accuracy alone cannot reveal this.

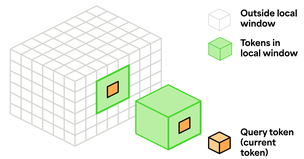
OUR FIX

PACK3D takes positions only from 3D coordinates

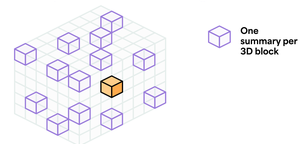
Its RoPE rotates each query and key by the token's 3D coordinate (z, y, x), never by its place in memory. The layer has three parts:

Sparse 3D Field View
Query at orange token

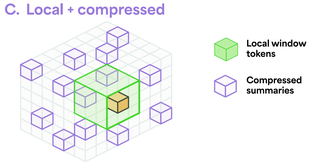
A. Local neighborhood (exact)



B. Compressed summaries (global)



C. Local + compressed



A Exact attention over each token's 3 × 3 × 3 neighborhood, found by coordinates, never by memory order

B Summaries pooled from token blocks, then from groups of blocks, cover the rest of the volume; keys are rotated before pooling

C Every query attends to both, in one layer

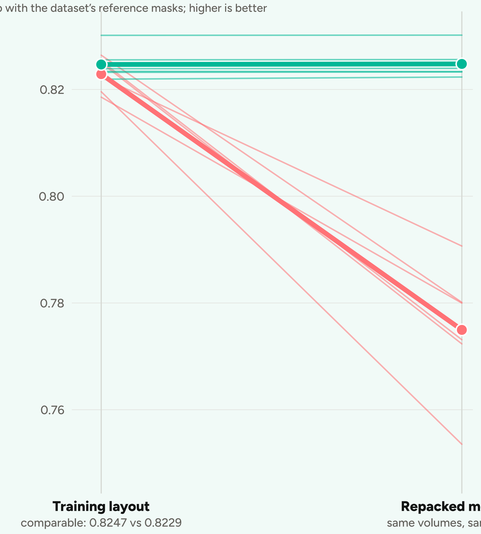
Reorder memory, carrying along everything that ties each token to the volume (its coordinate, neighbors, block and group, and slot in the block), and PACK3D's outputs move with the tokens, unchanged in exact arithmetic (Prop. 2).

THE TEST

Reorder the memory, keep the volume

Same trained models, same 120 validation volumes. Only the memory order changes: whole groups of token blocks move, once, the same way for every model, and predictions are put back in place before scoring. The control, sequence-index RoPE, is the same model with RoPE on memory position.

Segmentation Dice
overlap with the dataset's reference masks; higher is better



PACK3D
RoPE on 3D coordinates

unchanged
within 0.0005 in all 6 seeds
0.8247 to 0.8248

Sequence-index RoPE
same model, RoPE on memory position

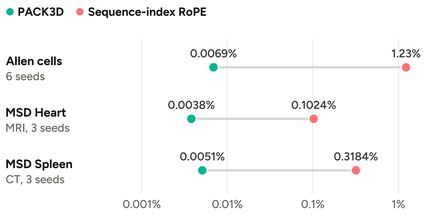
−0.048 Dice
lower in all 6 seeds
0.8229 to 0.7750

Allen cell images, 15 structures, 262,144 tokens each. Thin lines: 6 seeds; thick: mean.

Predicted labels that change	1.23% for sequence-index RoPE; 0.0069% for PACK3D, consistent with floating-point rounding
A different layout altogether	trained in raster blocks, tested in block-Morton order, which regroups the tokens that share a summary and so falls outside the guarantee (3 seeds): −0.0788 Dice for sequence-index RoPE, −0.0005 for PACK3D
Scope	one fixed reordering for every seed: a stress test, not an average over layouts

DOES IT HOLD UP?

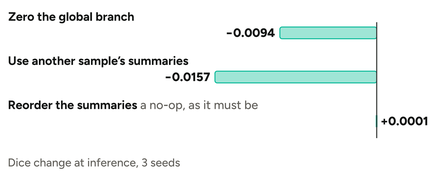
MRI and CT: layout dependence again, at little Dice cost



Predicted labels that change after the repack, mean of seeds (log scale)

On the Medical Segmentation Decathlon (MSD), Heart MRI and Spleen CT, sequence-index RoPE changes more predicted labels in every seed, yet mean Dice moves by less than 0.0003. Layout dependence can be present without costing much accuracy; its cost differs between settings.

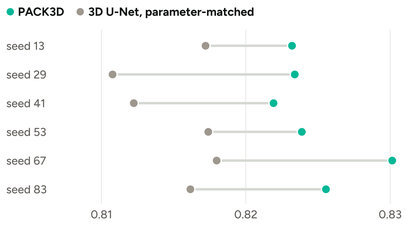
The trained model uses its global summaries



Dice change at inference, 3 seeds

Removing or swapping the summaries lowers Dice in all three seeds. In a separate synthetic probe, where only token positions inside a block differ, keys rotated before pooling reach 0.9997 accuracy; rotating at block centers, or not at all, stays at chance (0.5000).

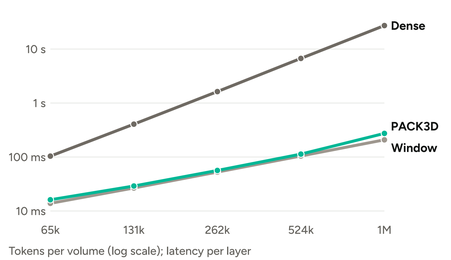
+0.0094 Dice over a matched 3D U-Net



Best-checkpoint Dice, Allen cells

Higher at the best checkpoint in all six seeds (paired 95% interval 0.0065 to 0.0123) and on 13 of 15 structures, with 30.08M vs 31.87M parameters.
A modest, matched comparison, not a ranking of attention above convolution.

Close to window attention in cost



Tokens per volume (log scale); latency per layer

One full layer, forward and backward, on one H200. Against dense attention (all token pairs): 6.43× faster at 65k tokens, 99.56× at 1M. Against window attention (neighbors only): 7–31% slower, with lower peak memory and an added global pathway.

What we do not claim, and a test you can run

Not claimed
that 3D RoPE is more accurate on a fixed layout; that these datasets need the global branch or keys rotated before pooling (a separately trained local-only model matches PACK3D's Dice); superiority over volumetric segmenters at large.

Scope
the target structure's own fluorescence channel is an input, so this is fluorescence-assisted segmentation, not label-free prediction from brightfield alone.

Try it on your model
permute the tokens in memory, carry along everything that ties each token to the volume, put the predictions back in spatial order, and compare.
A large change means they depend on storage order; near zero means this reordering left them unchanged.

```
p = torch.randperm(layout.geometry.num_tokens)
after = attn(x[:, p], layout.repack(p))[:, p.argsort()]
(after - attn(x, layout)).abs().max()  # ~0 for PACK3D
```

Data: Allen Institute for Cell Science, WTC-11 hiPSC Single-Cell Image Dataset v1, 15 tagged structures (AAVS1, ACTN1, ATP2A2, CETN2, DSP, FBL, HIST1H2BJ, LAMP1, LMNB1, MYH10, NPM1, NUP153, SEC61B, TOMM20, TUBA1B), with target fluorescence, brightfield and DNA as inputs; Medical Segmentation Decathlon, Heart and Spleen.
Code: github.com/amilworks/Pack3D · Supported by the Army Research Laboratory (W911NF-25-2-0159) and the National Science Foundation (DGE-2125644, 2411453).