SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer

Xuanyao Chen[1,2,#], Zhijian Liu[4,#], Haotian Tang[4],
Li Yi[1,3], Hang Zhao[1,3], Song Han[4]

[1]: Shanghai Qi Zhi Institute [2]: Fudan University [3]: Tsinghua University [4]: MIT

Background

High-resolution images have become ubiquitous!

Mobile Vision

iPhone 14
Pro Max

48MP

Mixed Reality

Apple
Vision Pro

23MP

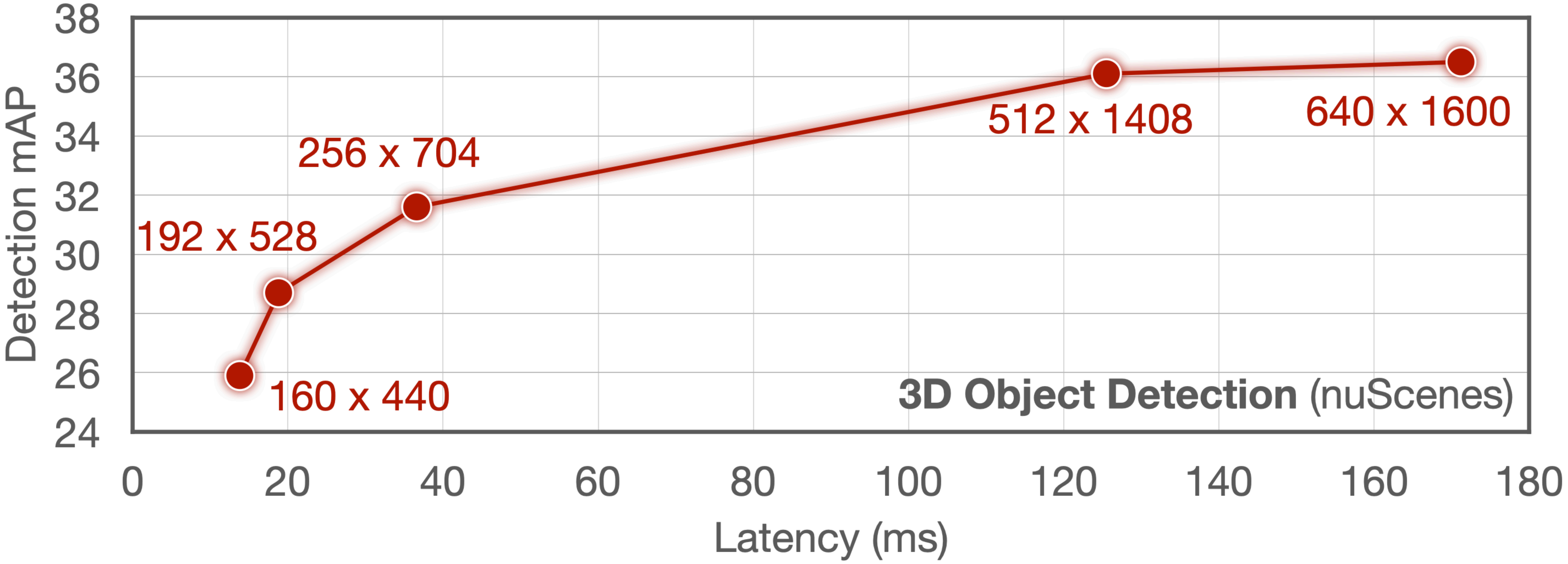
Autonomous Driving

Waymo
Driver

2MP x **29**

Background

Higher resolutions deliver better accuracy but also increase computation cost!

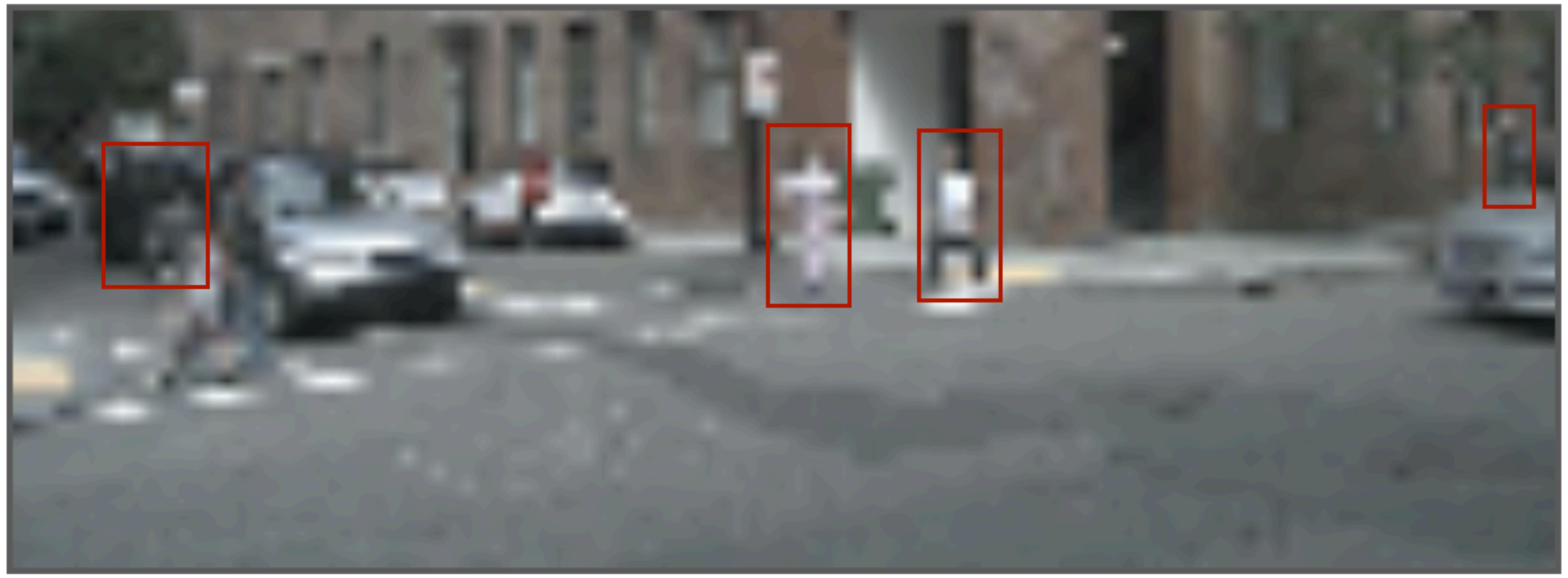


Activation Pruning

Sparse high-resolution features are better than dense low-resolution ones.

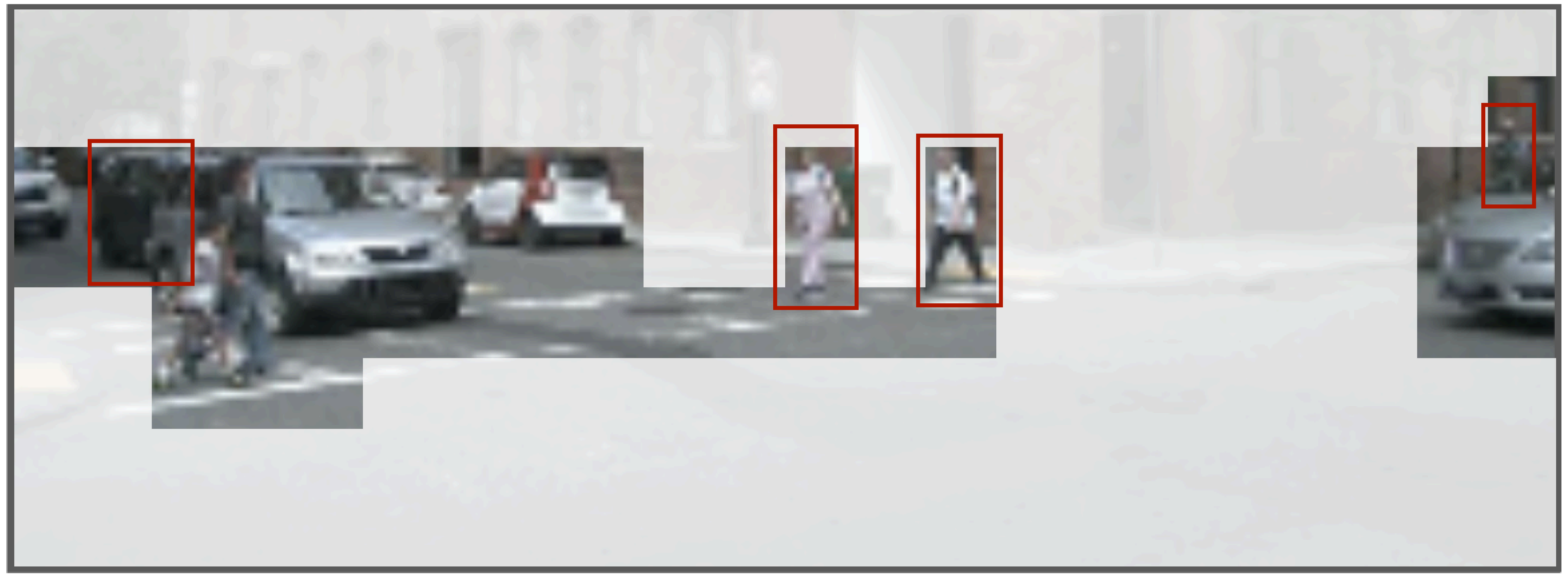
Uniform Resizing

Low Resolution (0.5X)
Dense Pixels (100%)



Activation Pruning

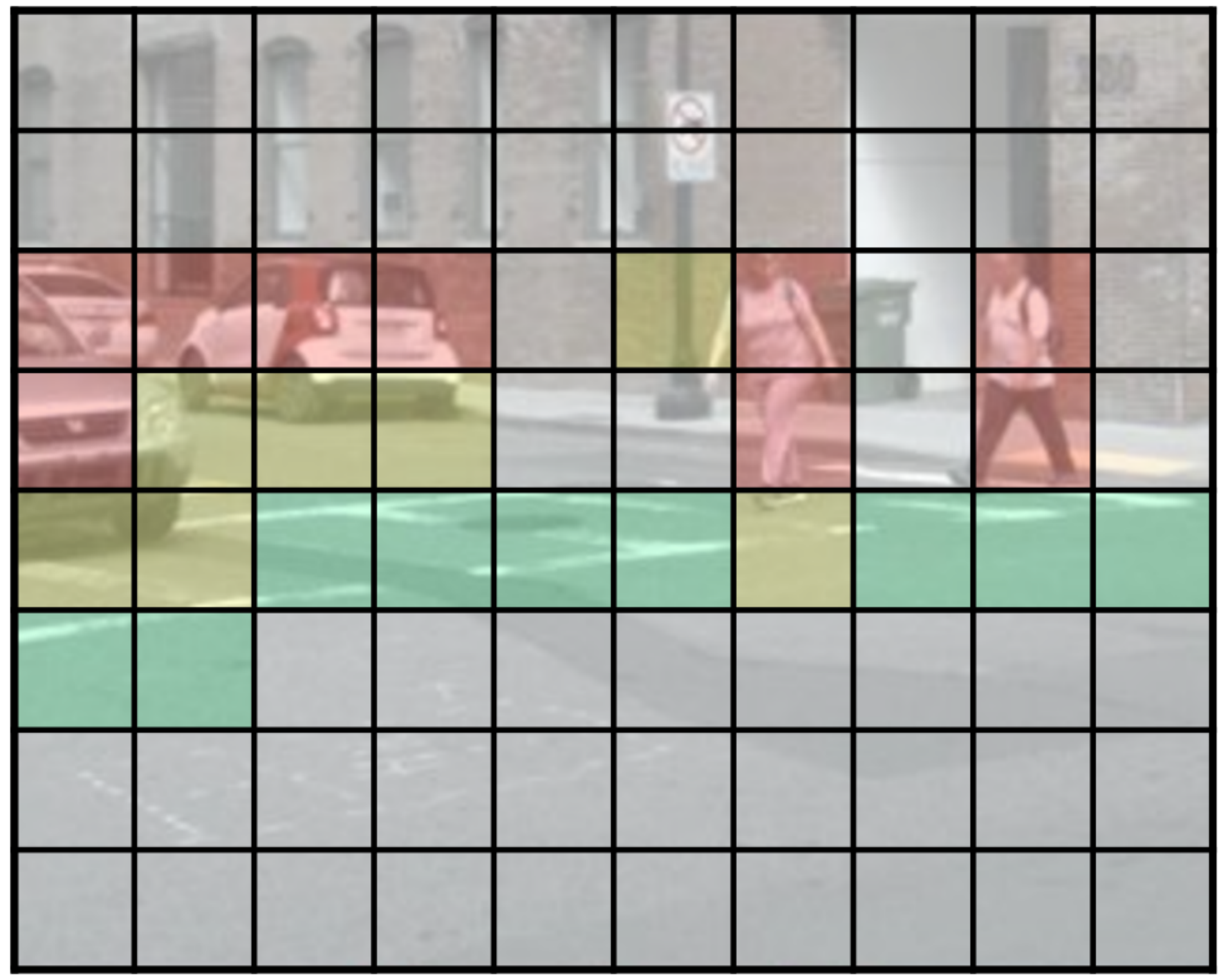
High Resolution (1X)
Sparse Pixels (25%)



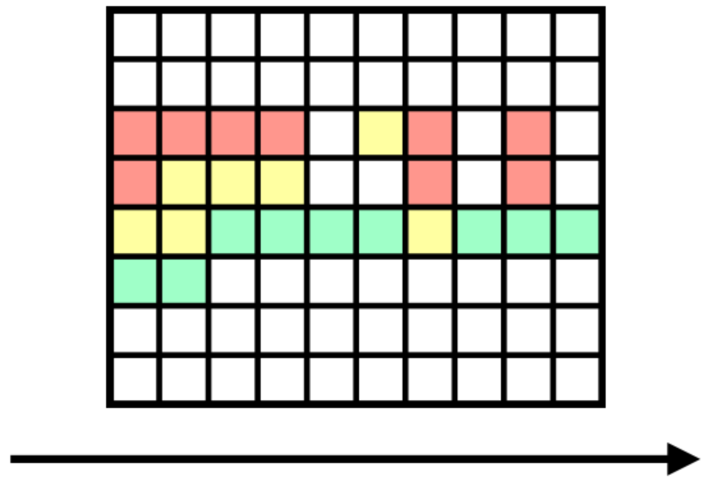
SparseViT — Sparse Vision Transformers

Step I. Window activation pruning (with non-uniform sparsity)

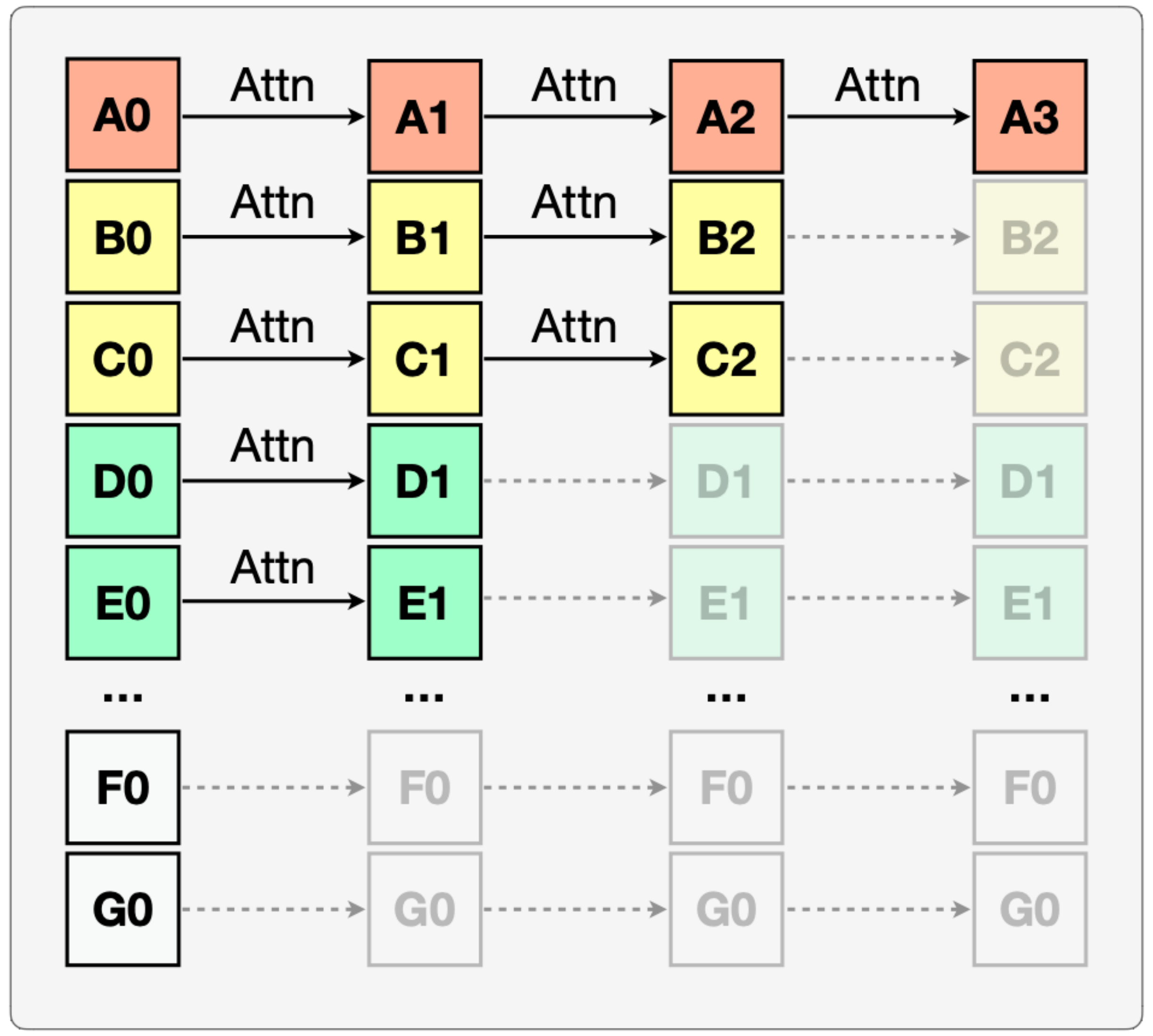
Input Image
(or Input Feature Map)



Window Importance
(L2 Activation Magnitude)



Sorting & Gathering

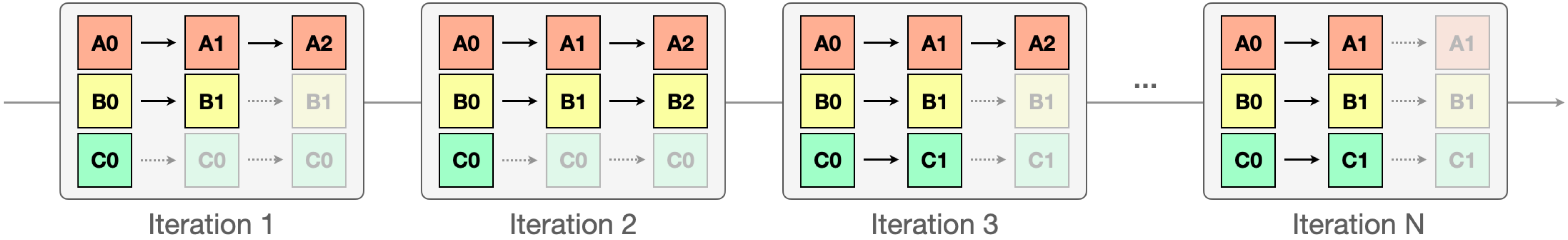


Sparse Window Attention

SparseViT — Sparse Vision Transformers

Step II. Sparsity-aware adaptation

Goal: Assess the model's accuracy under different activation sparsity settings both **efficiently** and **accurately**.

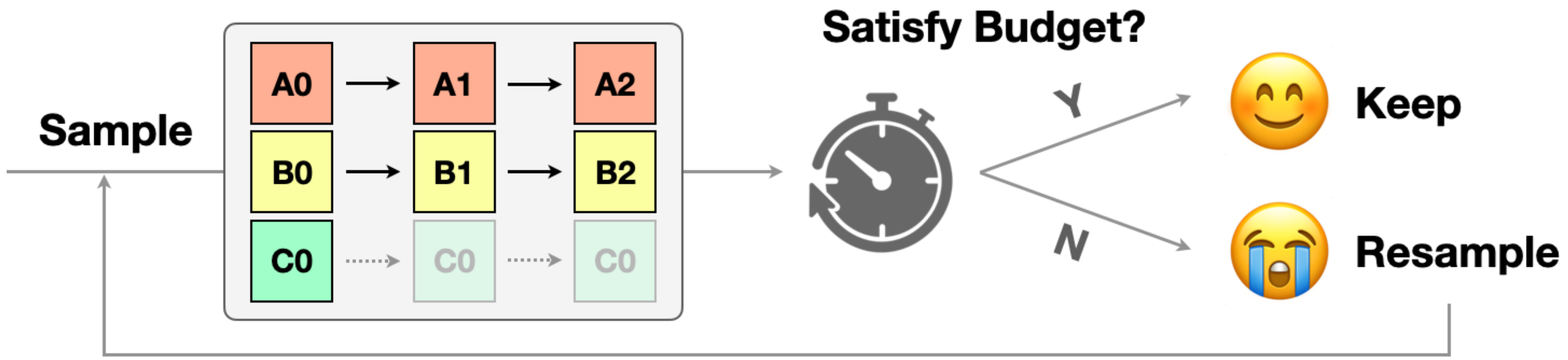


As the original model is trained with **only dense activations**, we **improve its sparsity awareness** by finetuning it with **randomly sampled** layerwise activation sparsity configurations at each training iteration.

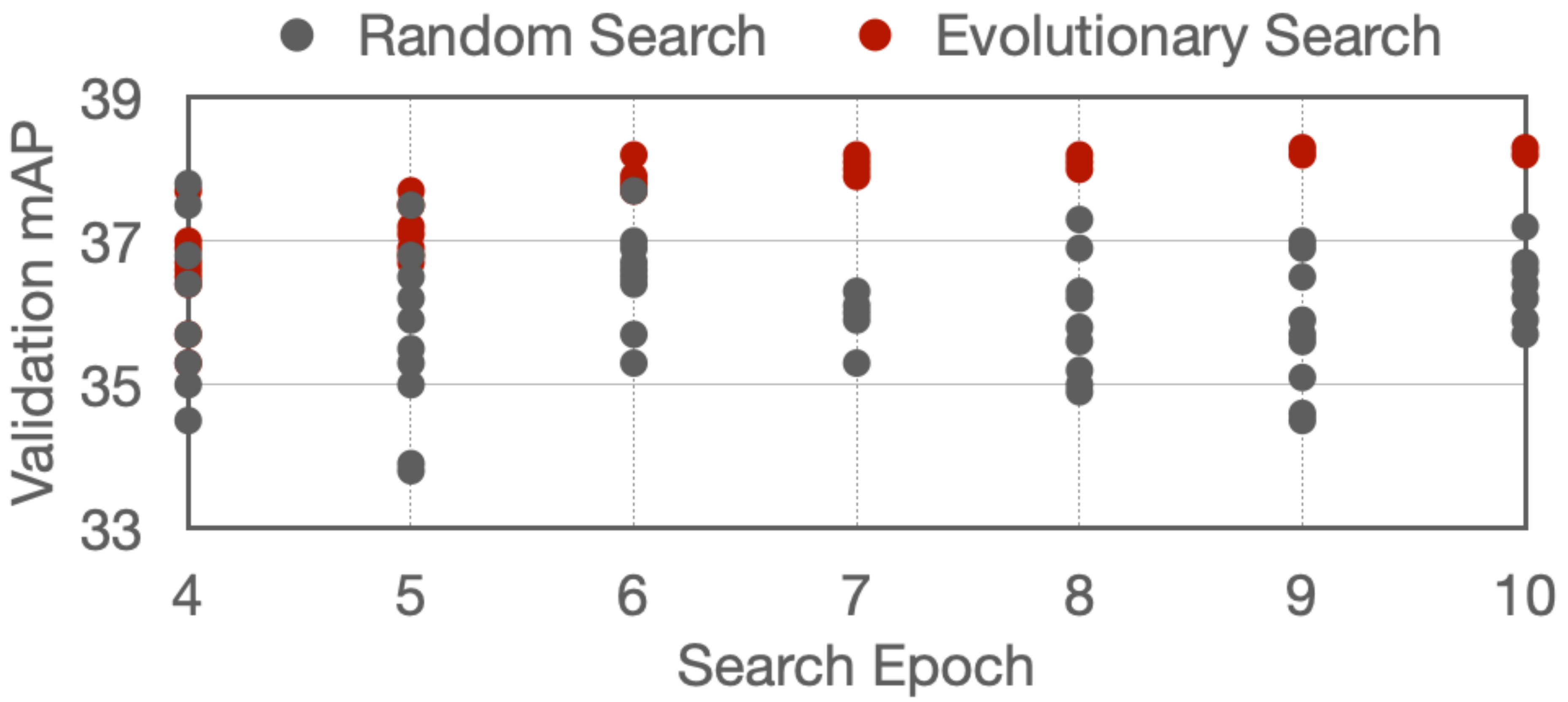
SparseViT — Sparse Vision Transformers

Step III. Resource-constrained search

Goal: Discover the **optimal** layerwise activation sparsity configuration **under a given latency budget**.

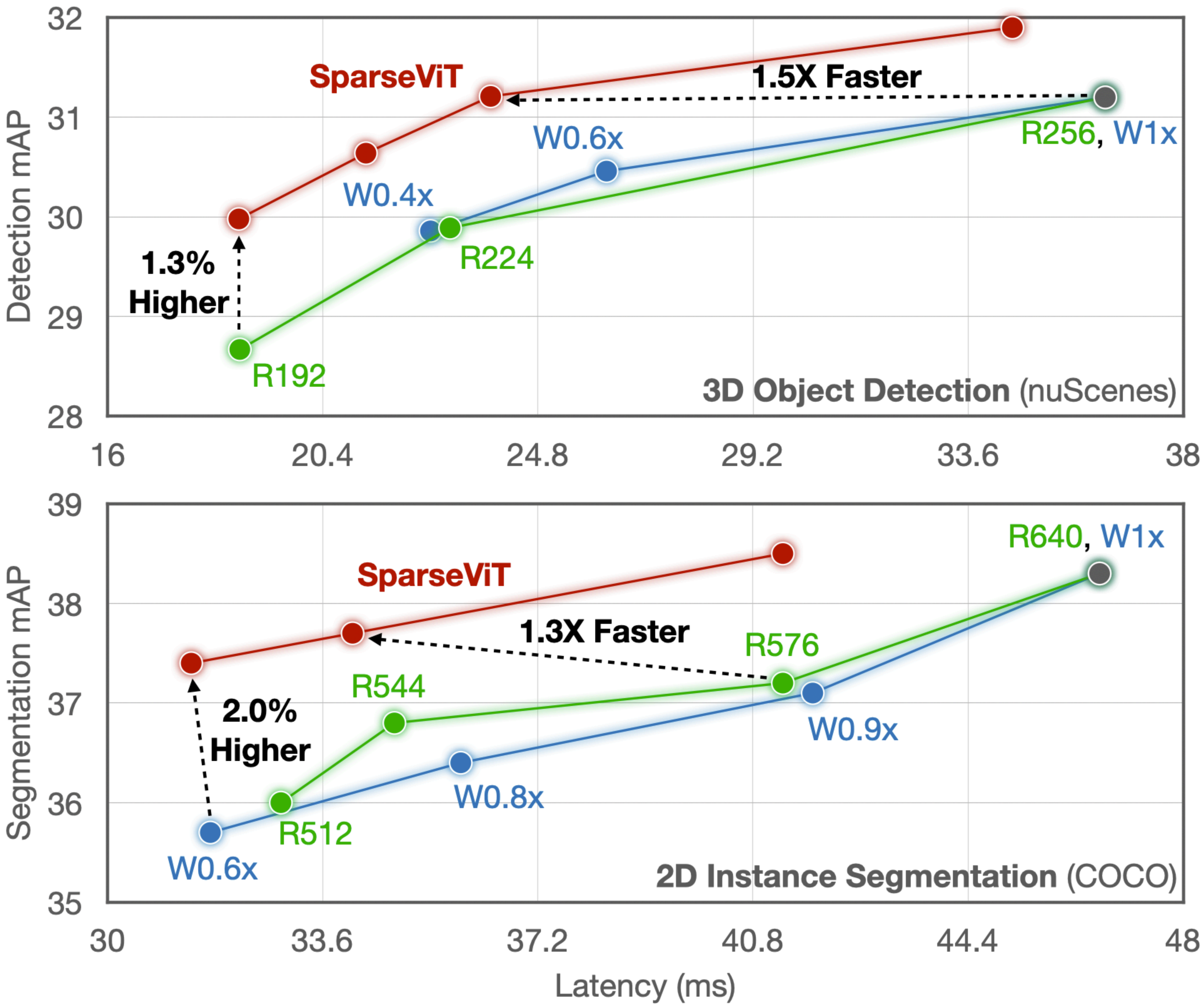


We enforce the latency constraint using **rejection sampling** (repeated resampling until satisfaction).



Evolutionary search is **sample-efficient**!

Results


Detection mAP
32
31
30
29
28
16
20.4
24.8
29.2
33.6
38
SparseViT
1.5X Faster
R256, W1x
W0.6x
W0.4x
R224
1.3%
Higher
R192
3D Object Detection (nuScenes)
Segmentation mAP
39
38
37
36
35
30
33.6
37.2
40.8
44.4
48
R640, W1x
SparseViT
1.3X Faster
R576
R544
2.0%
Higher
W0.9x
W0.8x
R512
W0.6x
2D Instance Segmentation (COCO)
Latency (ms)

Visualizations

SparseViT learns to prune **irrelevant background** windows while retaining **informative foreground** ones!