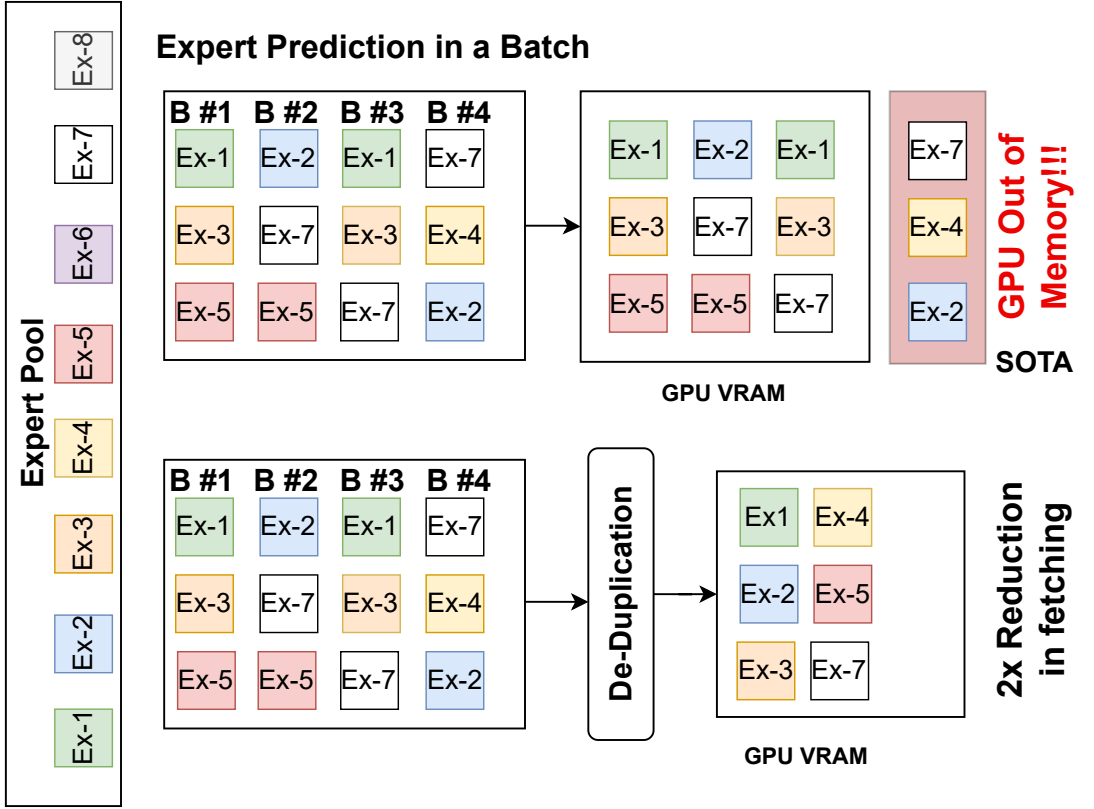
Expert Pool
Ex-8
Ex-7
Ex-6
Ex-5
Ex-4
Ex-3
Ex-2
Ex-1
Expert Prediction in a Batch
B #1
B #2
B #3
B #4
Ex-1
Ex-2
Ex-1
Ex-7
Ex-3
Ex-7
Ex-3
Ex-4
Ex-5
Ex-5
Ex-7
Ex-2
Ex-1
Ex-2
Ex-1
Ex-3
Ex-7
Ex-3
Ex-5
Ex-5
Ex-7
Ex-7
Ex-4
Ex-2
GPU Out of Memory!!!
SOTA
GPU VRAM
B #1
B #2
B #3
B #4
Ex-1
Ex-2
Ex-1
Ex-7
Ex-3
Ex-7
Ex-3
Ex-4
Ex-5
Ex-5
Ex-7
Ex-2
De-Duplication
Ex1
Ex-4
Ex-2
Ex-5
Ex-3
Ex-7
2x Reduction in fetching
GPU VRAM