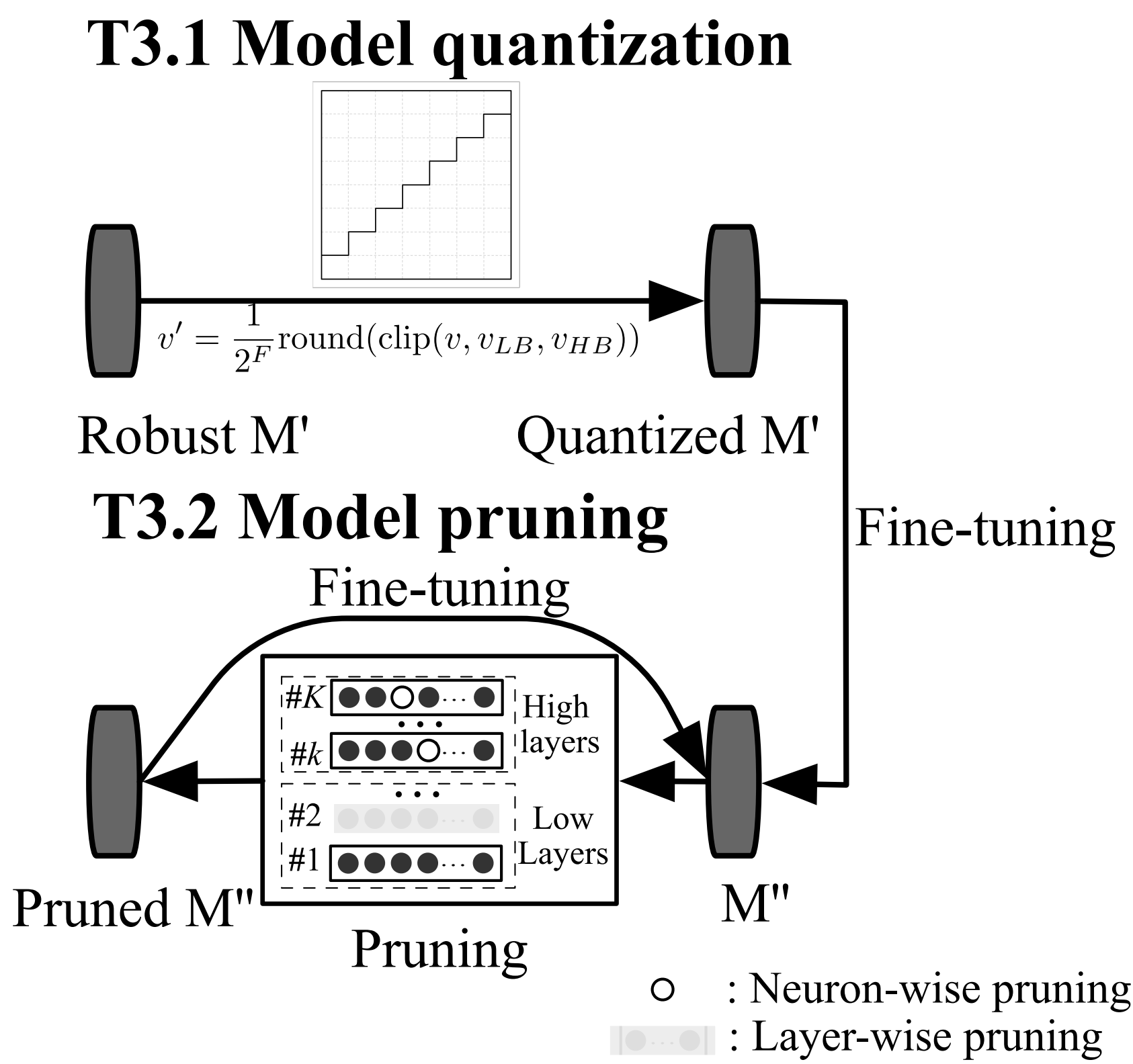
T3.1 Model quantization
$v' = \frac{1}{2^F}\text{round}(\text{clip}(v, v_{LB}, v_{HB}))$
Robust M'
Quantized M'
Fine-tuning
T3.2 Model pruning
Fine-tuning
#K
#k
High layers
#2
#1
Low Layers
Pruned M''
M''
Pruning
: Neuron-wise pruning
: Layer-wise pruning