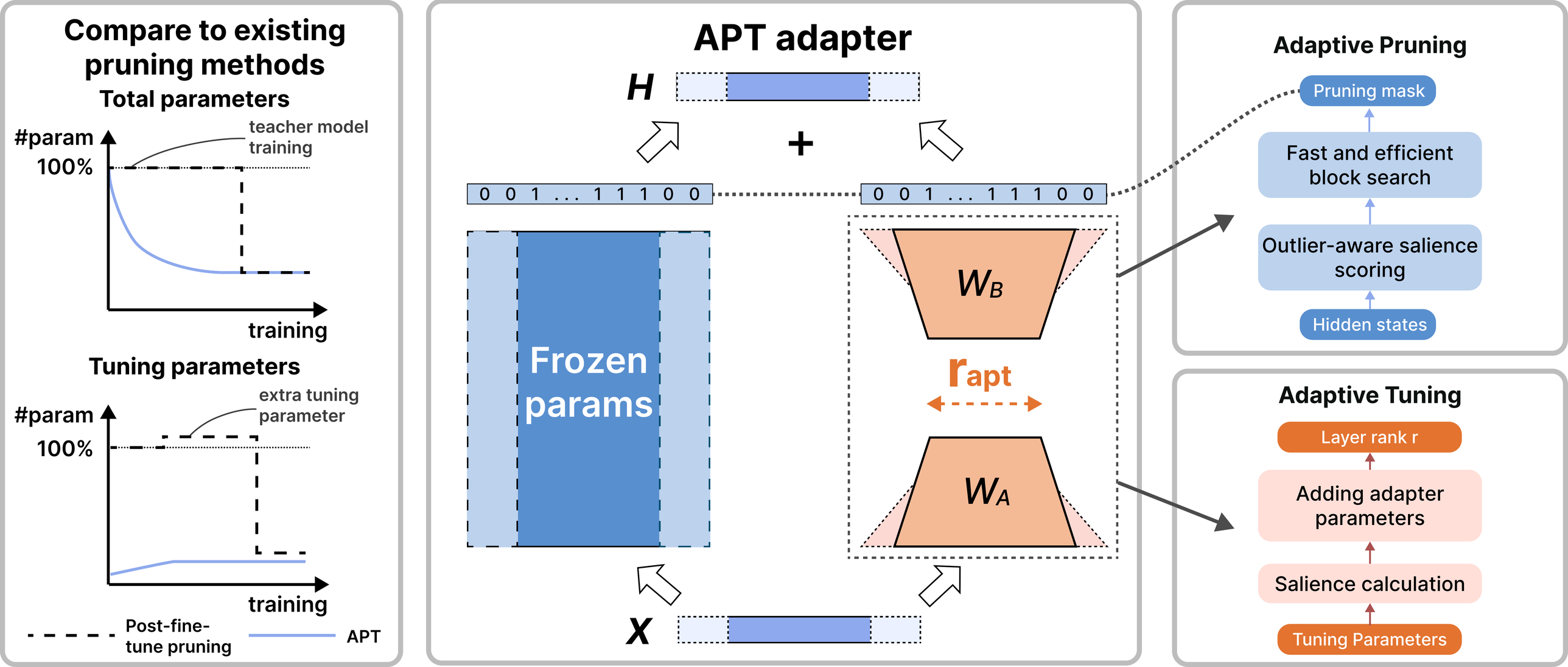
Compare to existing pruning methods
Total parameters
#param
100%
teacher model training
training
Tuning parameters
#param
100%
extra tuning parameter
training
Post-fine-tune pruning
APT
APT adapter
H
+
0 0 1 . . . 1 1 1 0 0
0 0 1 . . . 1 1 1 0 0
Frozen params
$W_B$
$r_{apt}$
$W_A$
X
Adaptive Pruning
Pruning mask
Fast and efficient block search
Outlier-aware salience scoring
Hidden states
Adaptive Tuning
Layer rank r
Adding adapter parameters
Salience calculation
Tuning Parameters