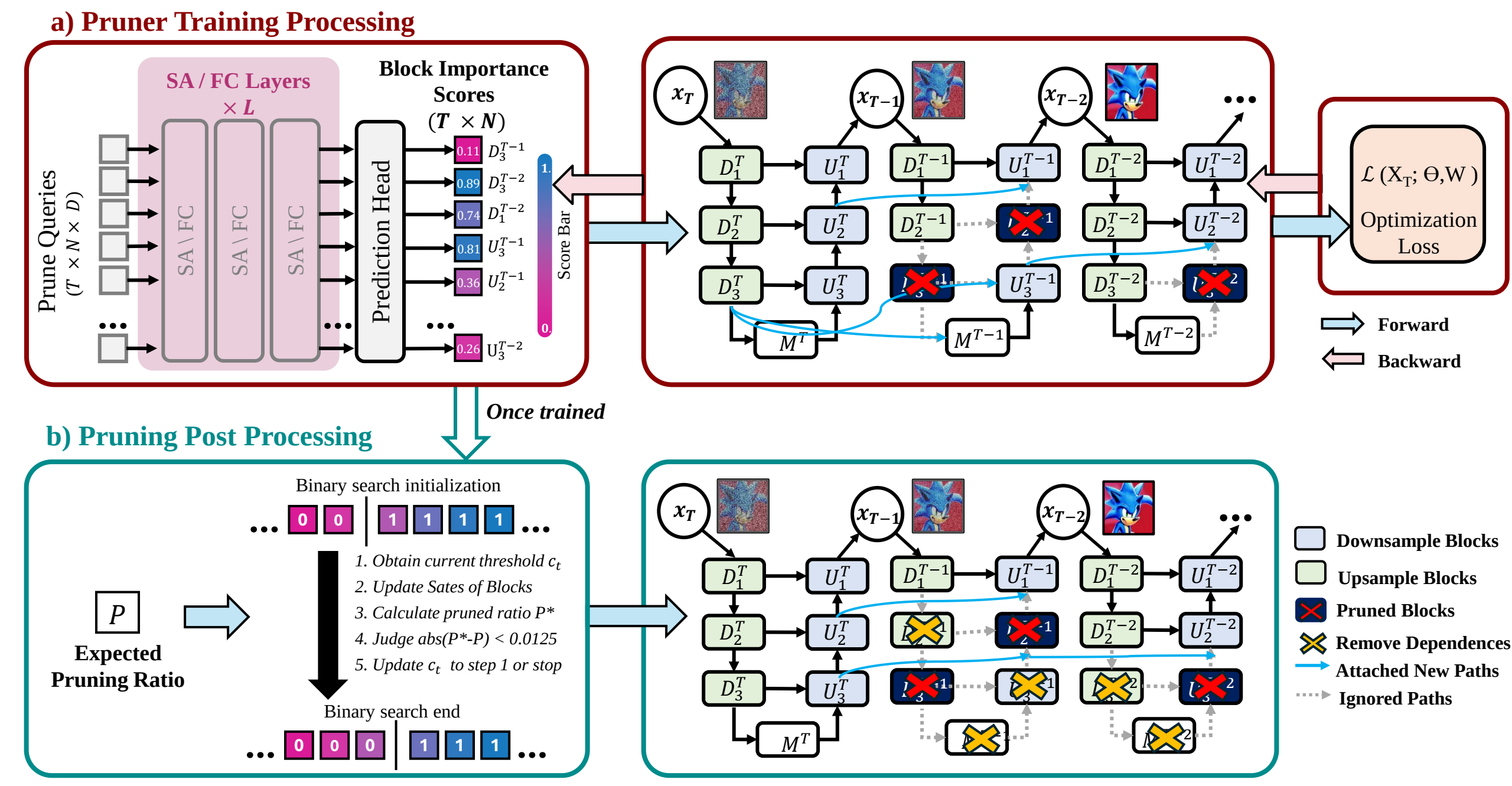
a) Pruner Training Processing
SA / FC Layers
× L
Prune Queries
(T × N × D)
SA \ FC
SA \ FC
SA \ FC
Prediction Head
Block Importance Scores
(T × N)
0.11
0.89
0.74
0.81
0.36
0.26
Score Bar
Once trained
b) Pruning Post Processing
Binary search initialization
0 0 1 1 1 1
1. Obtain current threshold c_t
2. Update Sates of Blocks
3. Calculate pruned ratio P*
4. Judge abs(P*-P) < 0.0125
5. Update c_t to step 1 or stop
Binary search end
0 0 0 1 1 1
P
Expected Pruning Ratio
$\mathcal{L}$ (X_T; Θ,W)
Optimization Loss
Forward
Backward
Downsample Blocks
Upsample Blocks
Pruned Blocks
Remove Dependences
Attached New Paths
Ignored Paths