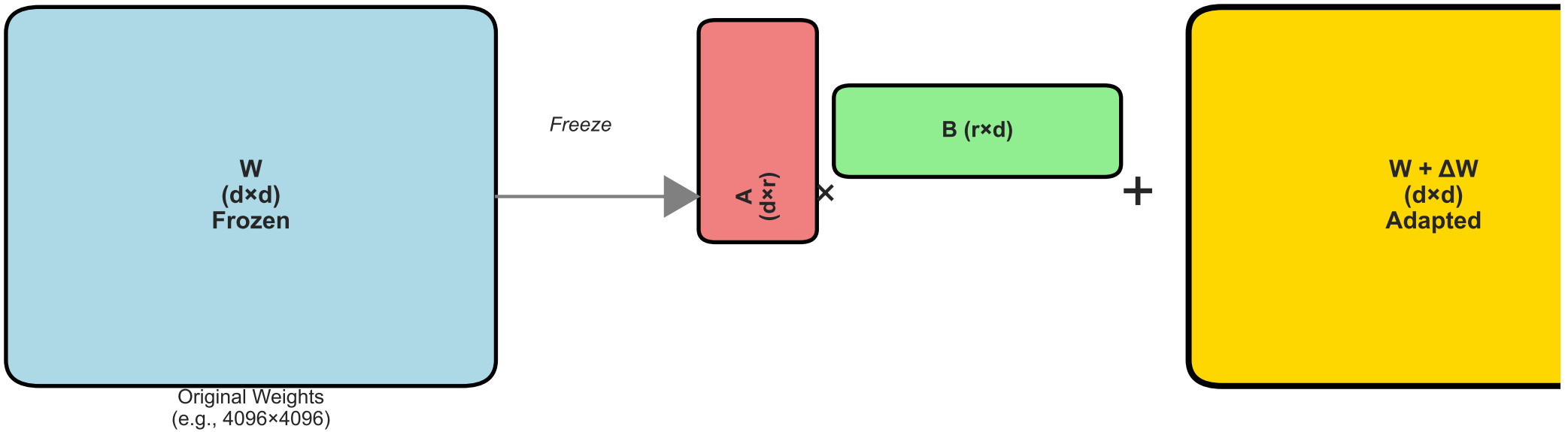
LoRA: Low-Rank Adaptation

Instead of updating 16M parameters, update only 32K!



Example: d=4096, r=8
Original: 4096×4096 = 16,777,216 parameters
LoRA: (4096×8) + (8×4096) = 65,536 parameters (0.39%!)