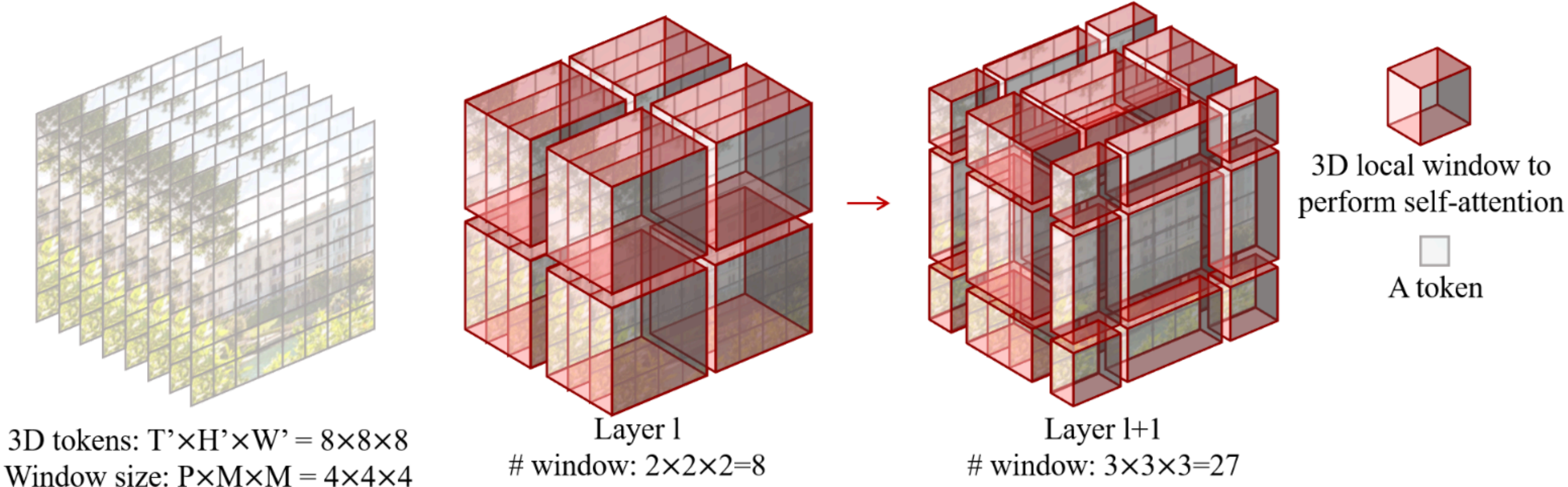
3D tokens: T’×H’×W’ = 8×8×8
Window size: P×M×M = 4×4×4
Layer l
# window: 2×2×2=8
Layer l+1
# window: 3×3×3=27
3D local window to perform self-attention
A token