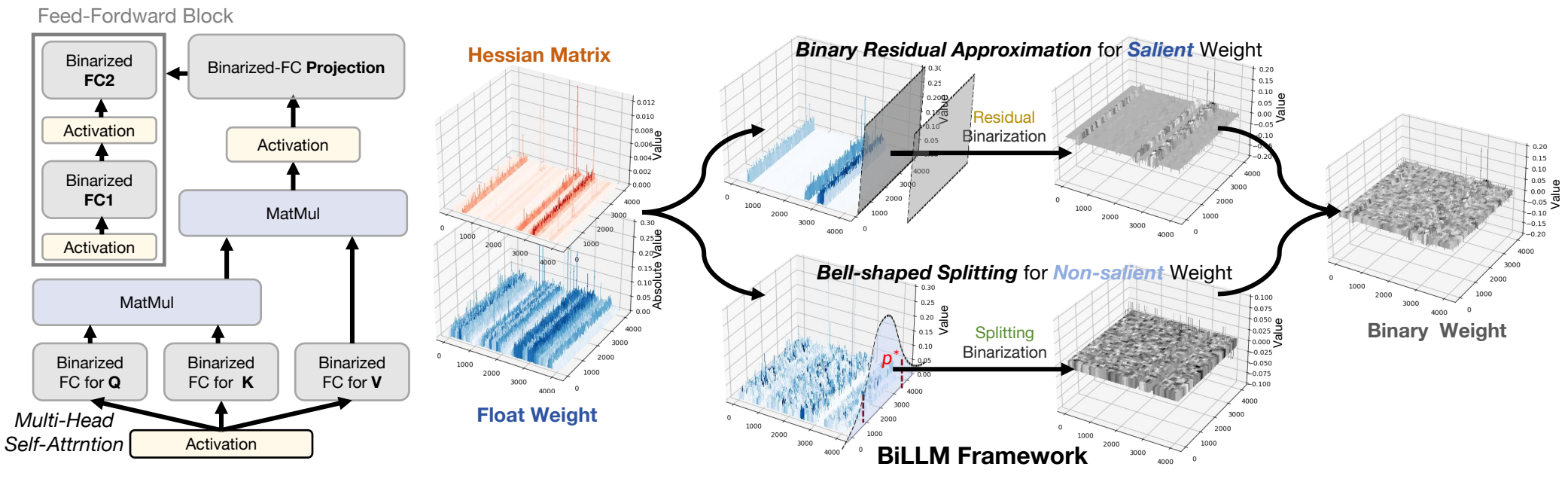
Feed-Fordward Block
Binarized FC2
Binarized-FC Projection
Activation
Activation
Binarized FC1
MatMul
Activation
MatMul
Binarized FC for Q
Binarized FC for K
Binarized FC for V
Multi-Head Self-Attrntion
Activation
Hessian Matrix
Float Weight
Binary Residual Approximation for Salient Weight
Residual Binarization
Bell-shaped Splitting for Non-salient Weight
Splitting Binarization
p*
BiLLM Framework
Binary Weight