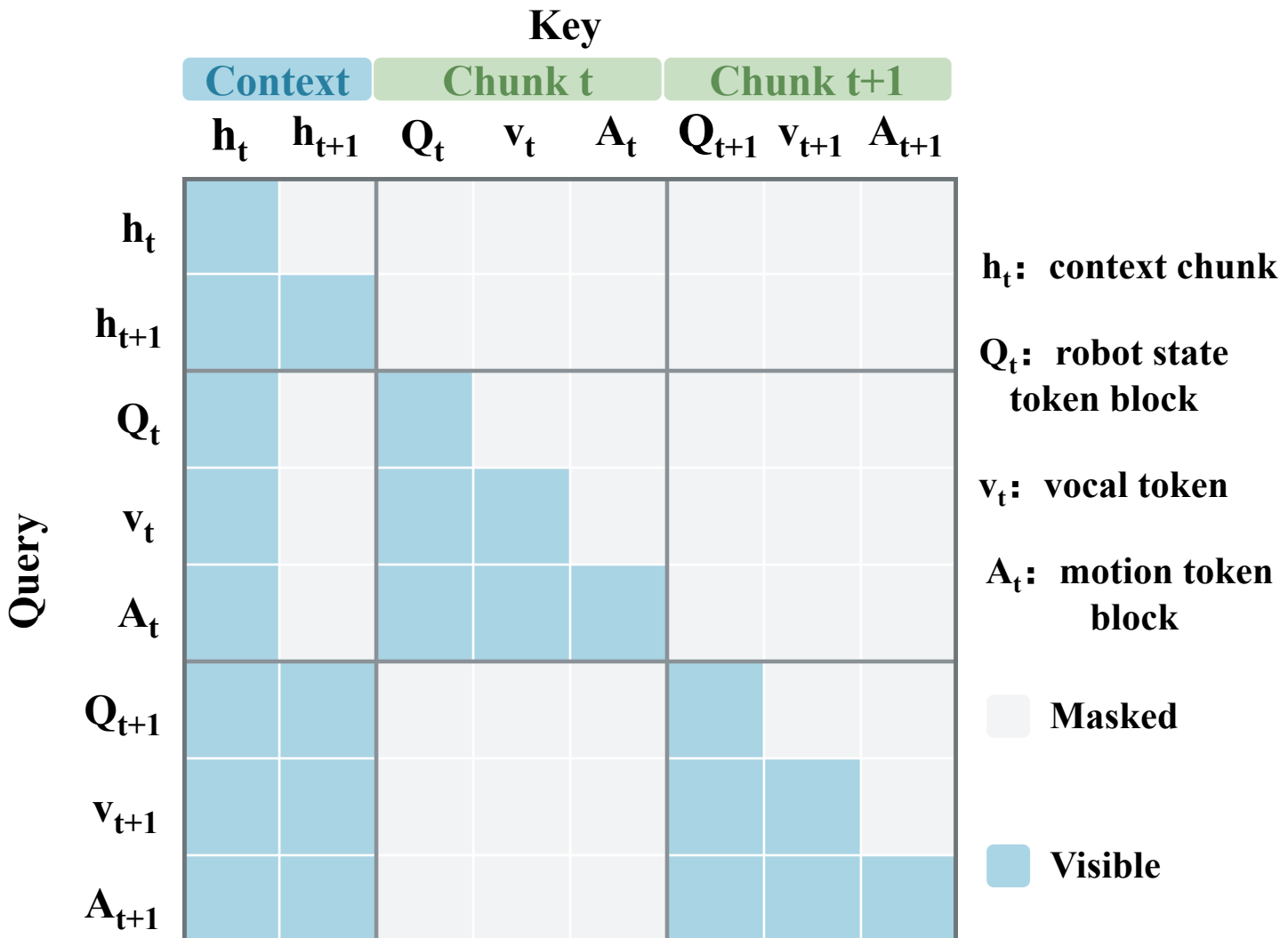

Key
Context
Chunk t
Chunk t+1
h_t h_{t+1} Q_t v_t A_t Q_{t+1} v_{t+1} A_{t+1}
Query
h_t
h_{t+1}
Q_t
v_t
A_t
Q_{t+1}
v_{t+1}
A_{t+1}
h_t: context chunk
Q_t: robot state token block
v_t: vocal token
A_t: motion token block
Masked
Visible