Input Sequence (L)

Value Sequence (L_V) Key Sequence (L_K)

Query Sequence (L_Q)

QProj

VSSM

KSSM

Values (V)

Keys (K)

Query (Q)
Single Token States

QK^T

Attention Mask (M)

$SM(QK^T + M)$

Values (V) Attention Scores (A)

VA

Next-Token Prediction