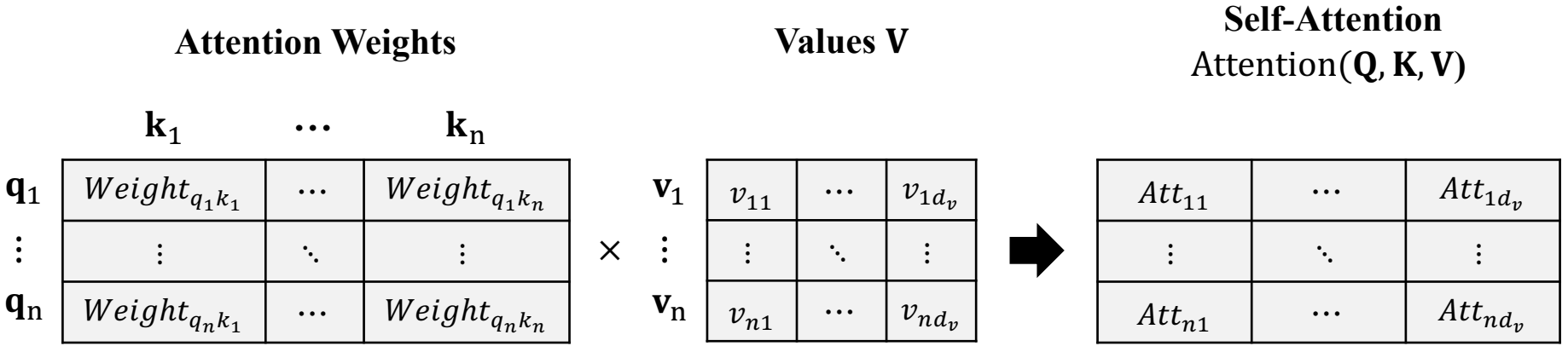
Attention Weights
Values V
Self-Attention
Attention(Q, K, V)
k1 ⋯ kn
q1
⋮
qn
Weight_{q1k1} ⋯ Weight_{q1kn}
⋮ ⋱ ⋮
Weight_{qnk1} ⋯ Weight_{qnkn}
×
v1
⋮
vn
v11 ⋯ v1dv
⋮ ⋱ ⋮
vn1 ⋯ vndv
Att11 ⋯ Att1dv
⋮ ⋱ ⋮
Attn1 ⋯ Attndv