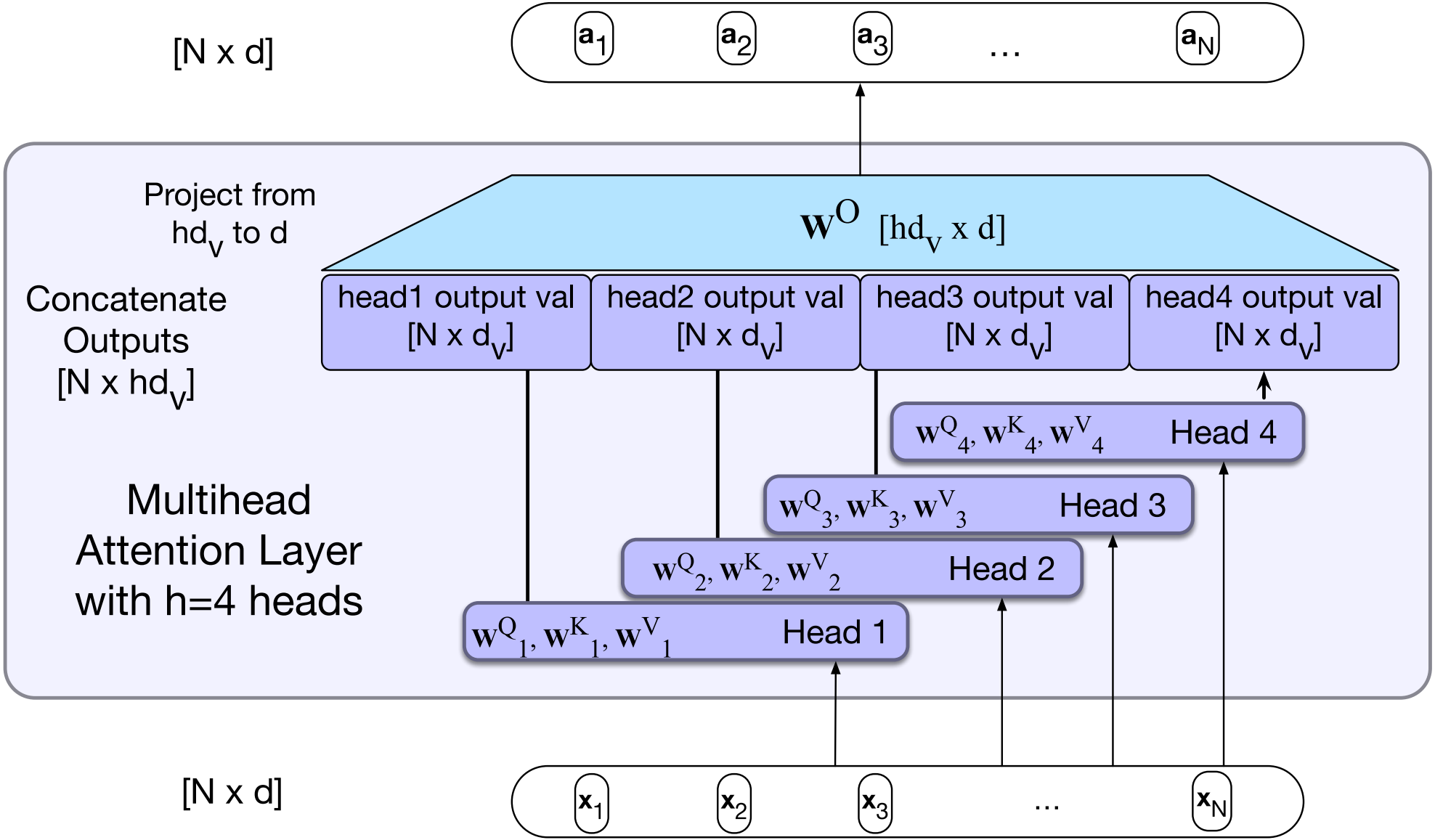
[N x d]
a1
a2
a3
…
aN
Project from hdV to d
W^O [hdV x d]
Concatenate Outputs [N x hdV]
head1 output val [N x dV]
head2 output val [N x dV]
head3 output val [N x dV]
head4 output val [N x dV]
Multihead Attention Layer with h=4 heads
W^Q4, W^K4, W^V4 Head 4
W^Q3, W^K3, W^V3 Head 3
W^Q2, W^K2, W^V2 Head 2
W^Q1, W^K1, W^V1 Head 1
[N x d]
x1
x2
x3
…
xN