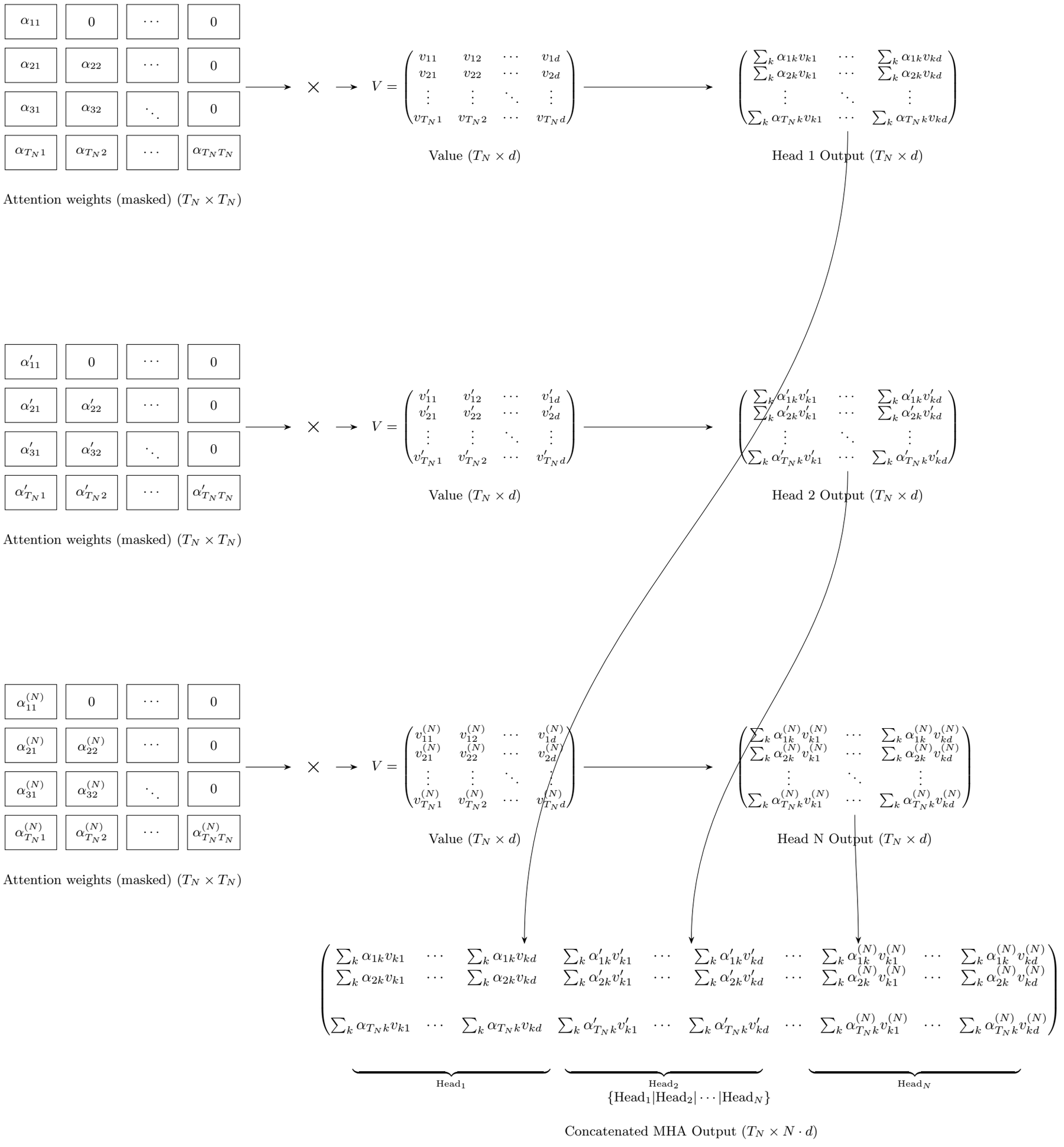
α_{11} 0 $\cdots$ 0
α_{21} α_{22} $\cdots$ 0
α_{31} α_{32} $\ddots$ 0
α_{T_N1} α_{T_N2} $\cdots$ $\alpha_{T_NT_N}$
Attention weights (masked) $(T_N \times T_N)$
$V =$
Value $(T_N \times d)$
$\sum_k \alpha_{1k} v_{k1}$ $\cdots$ $\sum_k \alpha_{1k} v_{kd}$
$\sum_k \alpha_{2k} v_{k1}$ $\cdots$ $\sum_k \alpha_{2k} v_{kd}$
$\sum_k \alpha_{T_Nk} v_{k1}$ $\cdots$ $\sum_k \alpha_{T_Nk} v_{kd}$
Head 1 Output $(T_N \times d)$
α'_{11} 0 $\cdots$ 0
α'_{21} α'_{22} $\cdots$ 0
α'_{31} α'_{32} $\ddots$ 0
α'_{T_N1} α'_{T_N2} $\cdots$ $\alpha'_{T_NT_N}$
Attention weights (masked) $(T_N \times T_N)$
$V =$
Value $(T_N \times d)$
$\sum_k \alpha'_{1k} v'_{k1}$ $\cdots$ $\sum_k \alpha'_{1k} v'_{kd}$
$\sum_k \alpha'_{2k} v'_{k1}$ $\cdots$ $\sum_k \alpha'_{2k} v'_{kd}$
$\sum_k \alpha'_{T_Nk} v'_{k1}$ $\cdots$ $\sum_k \alpha'_{T_Nk} v'_{kd}$
Head 2 Output $(T_N \times d)$
$\alpha^{(N)}_{11}$ 0 $\cdots$ 0
$\alpha^{(N)}_{21}$ $\alpha^{(N)}_{22}$ $\cdots$ 0
$\alpha^{(N)}_{31}$ $\alpha^{(N)}_{32}$ $\ddots$ 0
$\alpha^{(N)}_{T_N1}$ $\alpha^{(N)}_{T_N2}$ $\cdots$ $\alpha^{(N)}_{T_NT_N}$
Attention weights (masked) $(T_N \times T_N)$
$V =$
Value $(T_N \times d)$
$\sum_k \alpha^{(N)}_{1k} v^{(N)}_{k1}$ $\cdots$ $\sum_k \alpha^{(N)}_{1k} v^{(N)}_{kd}$
$\sum_k \alpha^{(N)}_{2k} v^{(N)}_{k1}$ $\cdots$ $\sum_k \alpha^{(N)}_{2k} v^{(N)}_{kd}$
$\sum_k \alpha^{(N)}_{T_Nk} v^{(N)}_{k1}$ $\cdots$ $\sum_k \alpha^{(N)}_{T_Nk} v^{(N)}_{kd}$
Head N Output $(T_N \times d)$
$\sum_k \alpha_{1k} v_{k1}$ $\cdots$ $\sum_k \alpha_{1k} v_{kd}$ $\sum_k \alpha'_{1k} v'_{k1}$ $\cdots$ $\sum_k \alpha'_{1k} v'_{kd}$ $\cdots$ $\sum_k \alpha^{(N)}_{1k} v^{(N)}_{k1}$ $\cdots$ $\sum_k \alpha^{(N)}_{1k} v^{(N)}_{kd}$
$\sum_k \alpha_{2k} v_{k1}$ $\cdots$ $\sum_k \alpha_{2k} v_{kd}$ $\sum_k \alpha'_{2k} v'_{k1}$ $\cdots$ $\sum_k \alpha'_{2k} v'_{kd}$ $\cdots$ $\sum_k \alpha^{(N)}_{2k} v^{(N)}_{k1}$ $\cdots$ $\sum_k \alpha^{(N)}_{2k} v^{(N)}_{kd}$
$\sum_k \alpha_{T_Nk} v_{k1}$ $\cdots$ $\sum_k \alpha_{T_Nk} v_{kd}$ $\sum_k \alpha'_{T_Nk} v'_{k1}$ $\cdots$ $\sum_k \alpha'_{T_Nk} v'_{kd}$ $\cdots$ $\sum_k \alpha^{(N)}_{T_Nk} v^{(N)}_{k1}$ $\cdots$ $\sum_k \alpha^{(N)}_{T_Nk} v^{(N)}_{kd}$
Head_1
Head_2
Head_N
$\{\text{Head}_1|\text{Head}_2|\cdots|\text{Head}_N\}$
Concatenated MHA Output $(T_N \times N \cdot d)$