$$\begin{aligned}
Input^e &= ([Input_{ohe}]^e_{n_W \times n_V} \bullet [W_{emb}]_{n_V \times n_E}) + P_e \\
Q_E &= Input^e \cdot Q_e; \quad K_E = Input^e \cdot K_e; \quad V_E = Input^e \cdot V_e \\
Ec_{t1} &= Norm(\sigma(\frac{Q_E K_E^T}{\sqrt{d_k}})V_E + Input^e) = Norm(A_e + Input^e) \\
FL_{e1} &= ReLu(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1}) \\
FL_{e2} &= FL_{e1} * W^e_{fl_2} + b^e_{fl_2} \\
Ec_{out} &= Norm(FL_{e2} + Ec_{t1})
\end{aligned}$$

$$\begin{aligned}
Input^d &= ([Input_{ohe}]^d_{n_W \times n_V} \bullet [W_{emb}]_{n_V \times n_E}) + P_e \\
Q_D &= Input^d \cdot Q_d \\
K_D &= Input^d \cdot K_d \\
V_D &= Input^d \cdot V_d \\
D_{t1} &= Norm(\sigma(Mask[\frac{Q_D K_D^T}{\sqrt{d_k}}])V_D + Input^d) = Norm(A_{mask} + Input^d) \\
Q_C &= D_{t1} \cdot Q_c \\
K_C &= Ec_{out} \cdot K_c \\
V_C &= Ec_{out} \cdot V_c \\
D_{t2} &= Norm\left(\sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right)V_C + D_{t1}\right) = Norm(A_{cr} + D_{t1}) \\
FL_{d1} &= ReLU(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1}) \\
FL_{d2} &= FL_{d1} \cdot W^d_{fl_2} + b^d_{fl_2} \\
D_{out} &= Norm(FL_{d2} + D_{t2}) = \frac{(FL_{d2} + D_{t2}) - \mu_{(FL_{d2}+D_{t2})}}{\sqrt{\sigma^2_{(FL_{d2}+D_{t2})} + \epsilon}} \\
Z_{out} &= D_{out} \cdot W^o + b^o \\
\sigma(Z^{out}) &= \sigma(D_{out} \cdot W^o + b^o)
\end{aligned}$$

$$A_e = \sigma(\frac{Q_E K_E^T}{\sqrt{d_k}})V_E$$

$$\begin{aligned}
\frac{\partial \text{Loss}}{\partial A_e} &= \frac{\partial \text{Loss}}{\partial Ec_{t1}} \cdot \frac{\partial Ec_{t1}}{\partial Norm(A_e + Input^e)} \cdot \frac{\partial Norm(A_e + Input^e)}{\partial A_e} \\
\frac{\partial \text{Loss}}{\partial Input^e_n} &= \frac{\partial \text{Loss}}{\partial Ec_{t1}} \cdot \frac{\partial Ec_{t1}}{\partial Norm(A_e + Input^e)} \cdot \frac{\partial Norm(A_e + Input^e)}{\partial Input^e} \\
\frac{\partial \text{Loss}}{\partial K_e} &= \frac{\partial \text{Loss}}{\partial A_e} \cdot \frac{\partial A_e}{\partial K_E} \cdot \frac{\partial K_E}{\partial K_e} = \frac{\partial \text{Loss}}{\partial A_e} \cdot \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right) \cdot \left[1 - \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right)\right] \cdot \frac{Q_E}{\sqrt{d_k}} \cdot V_E \cdot Input^e \\
\frac{\partial \text{Loss}}{\partial Q_e} &= \frac{\partial \text{Loss}}{\partial A_e} \cdot \frac{\partial A_e}{\partial Q_E} \cdot \frac{\partial Q_E}{\partial Q_e} = \frac{\partial \text{Loss}}{\partial A_e} \cdot \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right) \cdot \left[1 - \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right)\right] \cdot \frac{K_E^T}{\sqrt{d_k}} \cdot V_E \cdot Input^e \\
\frac{\partial \text{Loss}}{\partial V_e} &= \frac{\partial \text{Loss}}{\partial A_e} \cdot \frac{\partial A_e}{\partial V_E} \cdot \frac{\partial V_E}{\partial V_e} = \frac{\partial \text{Loss}}{\partial A_e} \cdot \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right) \cdot Input^e \\
\frac{\partial \text{Loss}}{\partial Input^e_Q} &= \frac{\partial \text{Loss}}{\partial A_e} \cdot \frac{\partial A_e}{\partial Q_E} \cdot \frac{\partial Q_E}{\partial Input^e} = \frac{\partial \text{Loss}}{\partial A_e} \cdot \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right) \cdot \left[1 - \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right)\right] \cdot \frac{K_E^T}{\sqrt{d_k}} \cdot V_E \cdot Q_e \\
\frac{\partial \text{Loss}}{\partial Input^e_K} &= \frac{\partial \text{Loss}}{\partial A_e} \cdot \frac{\partial A_e}{\partial K_E} \cdot \frac{\partial K_E}{\partial Input^e} = \frac{\partial \text{Loss}}{\partial A_e} \cdot \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right) \cdot \left[1 - \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right)\right] \cdot \frac{Q_E}{\sqrt{d_k}} \cdot V_E \cdot K_e \\
\frac{\partial \text{Loss}}{\partial Input^e_V} &= \frac{\partial \text{Loss}}{\partial A_e} \cdot \frac{\partial A_e}{\partial V_E} \cdot \frac{\partial V_E}{\partial Input^e} = \frac{\partial \text{Loss}}{\partial A_e} \cdot \sigma\left(\frac{Q_E K_E^T}{\sqrt{d_k}}\right) \cdot V_e \\
\frac{\partial \text{Loss}}{\partial Input^e} &= \frac{\partial \text{Loss}}{\partial Input^e_n} + \frac{\partial \text{Loss}}{\partial Input^e_V} + \frac{\partial \text{Loss}}{\partial Input^e_K} + \frac{\partial \text{Loss}}{\partial Input^e_Q} \\
\frac{\partial \text{Loss}}{\partial W^e_{emb}} &= \frac{\partial \text{Loss}}{\partial input^e} \cdot \frac{\partial input^e}{\partial W^e_{emb}} = \frac{\partial \text{Loss}}{\partial input^e} \cdot Input^e
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \text{Loss}}{\partial FL_{e2}} &= \frac{\partial \text{Loss}}{\partial Ec_{cout}} \cdot \frac{\partial Ec_{cout}}{\partial Norm(FL_{e2} + Ec_{t1})} \cdot \frac{\partial Norm(FL_{e2} + Ec_{t1})}{\partial FL_{e2}} \\
\frac{\partial \text{Loss}}{\partial E^n c_{t1}} &= \frac{\partial \text{Loss}}{\partial Ec_{cout}} \cdot \frac{\partial Ec_{cout}}{\partial Norm(FL_{e2} + Ec_{t1})} \cdot \frac{\partial Norm(FL_{e2} + Ec_{t1})}{\partial Ec_{t1}} \\
\frac{\partial \text{Loss}}{\partial W^e_{fl2}} &= \frac{\partial \text{Loss}}{\partial FL_{e2}} \cdot \frac{\partial FL_{e2}}{\partial W^e_{fl2}} = \frac{\partial \text{Loss}}{\partial FL_{e2}} \cdot Fl_{e1} \\
\frac{\partial \text{Loss}}{\partial b^e_{fl2}} &= \frac{\partial \text{Loss}}{\partial FL_{e2}} \cdot \frac{\partial FL_{e2}}{\partial b^e_{fl2}} = \frac{\partial \text{Loss}}{\partial FL_{e2}} \cdot 1 \\
\frac{\partial \text{Loss}}{\partial FL_{e1}} &= \frac{\partial \text{Loss}}{\partial FL_{e2}} \cdot \frac{\partial FL_{e2}}{\partial FL_{e1}} = \frac{\partial \text{Loss}}{\partial Fl_{e2}} \cdot W^e_{fl2} \\
\frac{\partial \text{Loss}}{\partial E_{ct1}} &= \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot \frac{\partial FL_{e1}}{\partial ReLu(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})} \cdot \frac{\partial ReLu(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})}{\partial(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})} \cdot \frac{\partial(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})}{\partial Ec_{t1}} = \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot \begin{cases} W^e_{fl_1}, & Ec_{t1} * W^e_{fl_1} + b^e_{fl_1} > 0 \\ 0, & \text{otherwise} \end{cases} \\
\frac{\partial \text{Loss}}{\partial W^e_{fl1}} &= \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot \frac{\partial FL_{e1}}{\partial ReLu(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})} \cdot \frac{\partial ReLu(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})}{\partial(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})} \cdot \frac{\partial(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})}{\partial W^e_{fl1}} = \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot \begin{cases} Ec_{t1}, & Ec_{t1} * W^e_{fl_1} + b^e_{fl_1} > 0 \\ 0, & \text{otherwise} \end{cases} \\
\frac{\partial \text{Loss}}{\partial b^e_{fl1}} &= \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot \frac{\partial FL_{e1}}{\partial ReLu(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})} \cdot \frac{\partial ReLu(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})}{\partial(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})} \cdot \frac{\partial(Ec_{t1} * W^e_{fl_1} + b^e_{fl_1})}{\partial b^e_{fl1}} = \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot \begin{cases} 1, & Ec_{t1} * W^e_{fl_1} + b^e_{fl_1} > 0 \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

$$A_{cr} = \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right) \cdot V_C$$

$$\begin{aligned}
\frac{\partial \text{Loss}}{\partial A_{cr}} &= \frac{\partial \text{Loss}}{\partial D_{t2}} \cdot \frac{\partial D_{t2}}{\partial Norm(A_{cr} + D_{t1})} \cdot \frac{\partial Norm(A_{cr} + D_{t1})}{\partial(A_{cr} + D_{t1})} \cdot \frac{\partial(A_{cr} + D_{t1})}{\partial A_{cr}} \\
\frac{\partial \text{Loss}}{\partial D_{t1}} &= \frac{\partial \text{Loss}}{\partial D_{t2}} \cdot \frac{\partial D_{t2}}{\partial Norm(A_{cr} + D_{t1})} \cdot \frac{\partial Norm(A_{cr} + D_{t1})}{\partial(A_{cr} + D_{t1})} \cdot \frac{\partial(A_{cr} + D_{t1})}{\partial D_{t1}} \\
\frac{\partial \text{Loss}}{\partial D^{qc}_{t1}} &= \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \frac{\partial A_{cr}}{\partial Q_C} \cdot \frac{\partial Q_C}{\partial D_{t1}} = \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right) \cdot \left[1 - \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right)\right] \cdot \frac{K_C^T}{\sqrt{d_k}} \cdot V_C \cdot Q_c \\
\frac{\partial \text{Loss}}{\partial Q_c} &= \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \frac{\partial A_{cr}}{\partial Q_C} \cdot \frac{\partial Q_C}{\partial Q_c} = \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right) \cdot \left[1 - \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right)\right] \cdot \frac{K_C^T}{\sqrt{d_k}} \cdot V_C \cdot D_{t1} \\
\frac{\partial \text{Loss}}{\partial K_c} &= \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \frac{\partial A_{cr}}{\partial K_C} \cdot \frac{\partial K_C}{\partial K_c} = \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right) \cdot \left[1 - \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right)\right] \cdot \frac{Q_C}{\sqrt{d_k}} \cdot V_C \cdot Ec_{out} \\
\frac{\partial \text{Loss}}{\partial V_c} &= \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \frac{\partial A_{cr}}{\partial V_C} \cdot \frac{\partial V_C}{\partial V_c} = \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right) \cdot Ec_{out} \\
\frac{\partial \text{Loss}}{\partial Ec^K_{cout}} &= \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \frac{\partial A_{cr}}{\partial K_C} \cdot \frac{\partial K_C}{\partial Ec_{cout}} = \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right) \cdot \left[1 - \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right)\right] \cdot \frac{Q_C}{\sqrt{d_k}} \cdot V_C \cdot K_c \\
\frac{\partial \text{Loss}}{\partial Ec^V_{cout}} &= \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \frac{\partial A_{cr}}{\partial V_C} \cdot \frac{\partial V_C}{\partial Ec_{cout}} = \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \sigma\left(\frac{Q_C K_C^T}{\sqrt{d_k}}\right) \cdot V_c \\
\frac{\partial \text{Loss}}{\partial Ec_{cout}} &= \frac{\partial \text{Loss}}{\partial Ec^K_{cout}} + \frac{\partial \text{Loss}}{\partial Ec^V_{cout}}
\end{aligned}$$

$$A_{mask} = \sigma(Mask[\frac{Q_D K_D^T}{\sqrt{d_k}}])V_D = \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask)V_D$$

$$\begin{aligned}
\frac{\partial \text{Loss}}{\partial A_{mask}} &= \frac{\partial \text{Loss}}{\partial D_{t1}} \cdot \frac{\partial D_{t1}}{\partial Norm(A_{mask} + Input^d)} \cdot \frac{\partial Norm(A_{mask} + Input^d)}{\partial A_{mask}} \\
\frac{\partial \text{Loss}}{\partial input^d_{dt1}} &= \frac{\partial \text{Loss}}{\partial D_{t1}} \cdot \frac{\partial D_{t1}}{\partial Norm(A_{mask} + Input^d)} \cdot \frac{\partial Norm(A_{mask} + Input^d)}{\partial input^d} \\
\frac{\partial \text{Loss}}{\partial Q_d} &= \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \frac{\partial A_{mask}}{\partial Q_D} \cdot \frac{\partial Q_D}{\partial Q_d} = \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask) \cdot \left[1 - \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask)\right] \cdot \frac{K_D^T}{\sqrt{d_k}} \cdot V_D \cdot Input^d \\
\frac{\partial \text{Loss}}{\partial K_d} &= \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \frac{\partial A_{mask}}{\partial K_D} \cdot \frac{\partial K_D}{\partial K_d} = \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask) \cdot \left[1 - \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask)\right] \cdot \frac{Q_D}{\sqrt{d_k}} \cdot V_D \cdot Input^d \\
\frac{\partial \text{Loss}}{\partial V_d} &= \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \frac{\partial A_{mask}}{\partial V_D} \cdot \frac{\partial V_D}{\partial V_d} = \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask) \cdot Input^d \\
\frac{\partial \text{Loss}}{\partial input^q_d} &= \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \frac{\partial A_{mask}}{\partial Q_D} \cdot \frac{\partial Q_D}{\partial input_d} = \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask) \cdot \left[1 - \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask)\right] \cdot \frac{K_D^T}{\sqrt{d_k}} \cdot V_D \cdot Q_d \\
\frac{\partial \text{Loss}}{\partial input^k_d} &= \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \frac{\partial A_{mask}}{\partial K_D} \cdot \frac{\partial K_D}{\partial input_d} = \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask) \cdot \left[1 - \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask)\right] \cdot \frac{Q_D}{\sqrt{d_k}} \cdot V_D \cdot K_d \\
\frac{\partial \text{Loss}}{\partial input^v_d} &= \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \frac{\partial A_{mask}}{\partial V_D} \cdot \frac{\partial V_D}{\partial input_d} = \frac{\partial \text{Loss}}{\partial A_{mask}} \cdot \sigma(\frac{Q_D K_D^T}{\sqrt{d_k}} + Mask) \cdot V_d \\
\frac{\partial \text{Loss}}{\partial input_d} &= \frac{\partial \text{Loss}}{\partial input^q_d} + \frac{\partial \text{Loss}}{\partial input^k_d} + \frac{\partial \text{Loss}}{\partial input^v_d} + \frac{\partial \text{Loss}}{\partial input^d_{dt1}} \\
\frac{\partial \text{Loss}}{\partial W_{emb}} &= \frac{\partial \text{Loss}}{\partial input_d} \cdot \frac{\partial input_d}{\partial W_{emb}} = \frac{\partial \text{Loss}}{\partial input_d} \cdot Input_d
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \text{Loss}}{\partial Z^{out}} &= \frac{\partial \text{Loss}}{\partial \sigma(Z^{out})} \cdot \frac{\partial \sigma(\partial Z^{out})}{\partial Z^{out}} = \sigma(Z^{out}_i) - y_i \\
\frac{\partial \text{Loss}}{\partial D_{out}} &= \frac{\partial \text{Loss}}{\partial \sigma(Z^{out})} \cdot \frac{\partial \sigma(\partial Z^{out})}{\partial Z^{out}} \cdot \frac{\partial Z^{out}}{\partial D_{out}} = (\sigma(Z^{out}_i) - y_i) \cdot W^o \\
\frac{\partial \text{Loss}}{\partial FL_{d2}} &= \frac{\partial \text{Loss}}{\partial \sigma(Z^{out})} \cdot \frac{\partial \sigma(\partial Z^{out})}{\partial Z^{out}} \cdot \frac{\partial Z^{out}}{\partial D_{out}} \cdot \frac{\partial D_{out}}{\partial Norm(FL_{d2} + D_{t2})} \cdot \frac{\partial Norm(FL_{d2} + D_{t2})}{\partial FL_{d2}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \text{Loss}}{\partial FL_{d1}} &= \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot \frac{\partial FL_{d2}}{\partial FL_{d1}} = \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot W^d_{fl2} \\
\frac{\partial \text{Loss}}{\partial W^d_{fl2}} &= \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot \frac{\partial FL_{d2}}{\partial W^d_{fl2}} = \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot FL_{d1} \\
\frac{\partial \text{Loss}}{\partial b^d_{fl2}} &= \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot \frac{\partial FL_{d2}}{\partial b^d_{fl2}} = \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot 1 \\
\frac{\partial \text{Loss}}{\partial D_{t2}} &= \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot \frac{\partial FL_{d1}}{\partial ReLu(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})} \cdot \frac{\partial ReLu(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})}{\partial(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})} \cdot \frac{\partial(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})}{\partial D_{t2}} = \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot \begin{cases} W^d_{fl_1}, & D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1} > 0 \\ 0, & \text{otherwise} \end{cases} \\
\frac{\partial \text{Loss}}{\partial W_{fl1}} &= \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot \frac{\partial FL_{d1}}{\partial ReLu(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})} \cdot \frac{\partial ReLu(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})}{\partial(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})} \cdot \frac{\partial(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})}{\partial W_{fl1}} = \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot \begin{cases} D_{t2}, & D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1} > 0 \\ 0, & \text{otherwise} \end{cases} \\
\frac{\partial \text{Loss}}{\partial b_{fl1}} &= \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot \frac{\partial FL_{d1}}{\partial ReLu(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})} \cdot \frac{\partial ReLu(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})}{\partial(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})} \cdot \frac{\partial(D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1})}{\partial b_{fl1}} = \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot \begin{cases} 1, & D_{t2} \cdot W^d_{fl_1} + b^d_{fl_1} > 0 \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

$$\frac{\partial D_{out}}{\partial FL_{d2}} = \frac{\partial Norm(FL_{d2} + D_{t2})}{\partial FL_{d2}} = \frac{\partial \frac{(FL_{d2}+D_{t2}) - \mu_{(FL_{d2}+D_{t2})}}{\sqrt{\sigma^2_{(FL_{d2}+D_{t2})} + \epsilon}}}{\partial FL_{d2}}$$

$$X = FL_{d2} + D_{t2};$$

$$\begin{aligned}
\frac{\partial D_{out}}{\partial FL_{d2}} &= \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial X} \cdot \frac{\partial X}{\partial FL_{d2}} \\
\frac{\partial D_{out}}{\partial D_{t2}} &= \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial X} \cdot \frac{\partial X}{\partial D_{t2}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial X} &= \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial X} + \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial \mu_{(X)}} \cdot \frac{\partial \mu_{(X)}}{\partial X} + \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial \sigma^2_{(X)}} \cdot \frac{\partial \sigma^2_{(X)}}{\partial X} \\
\frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial X} &= \frac{1}{\sqrt{\sigma^2_{(X)} + \epsilon}}
\end{aligned}$$

$$\frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial \mu_{(X)}} \cdot \frac{\partial \mu_{(X)}}{\partial X} = -\frac{1}{\sqrt{\sigma^2_{(X)} + \epsilon}} \cdot \frac{1}{N}$$

$$\sigma^2_{(X)} = \frac{1}{N} \sum (X - \mu_x)^2; \quad \frac{\partial \sigma^2_{(X)}}{\partial X} = \frac{2}{N}(X - \mu_x) \cdot 1$$

$$\frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma^2_{(X)} + \epsilon}}}{\partial \sigma^2_{(X)}} \cdot \frac{\partial \sigma^2_{(X)}}{\partial X} = \frac{\partial}{\partial \sigma^2_{(X)}} \frac{X - \mu_{(X)}}{(\sigma^2_{(X)} + \epsilon)^{-1/2}} \cdot \frac{\partial \sigma^2_{(X)}}{\partial X} = -\frac{1}{2}(\sigma^2_{(X)} + \epsilon)^{-3/2} \cdot 1 \cdot (X - \mu_{(X)}) \frac{\partial \sigma^2_{(X)}}{\partial X} = -\frac{1}{2} \frac{X - \mu_{(X)}}{(\sigma^2_{(X)} + \epsilon)^{3/2}} \cdot \frac{2}{N}(X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma^2_{(X)} + \epsilon)^{3/2}}$$

$$\begin{aligned}
\frac{\partial D_{out}}{\partial FL_{d2}} &= (1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma^2_{(X)} + \epsilon}} - \frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma^2_{(X)} + \epsilon)^{3/2}} = \frac{\partial D_{out}}{\partial D_{t2}} \\
\frac{\partial \text{Loss}}{\partial FL_{d2}} &= \frac{\partial \text{Loss}}{\partial \sigma(Z^{out})} \cdot \frac{\partial \sigma(\partial Z^{out})}{\partial Z^{out}} \cdot \frac{\partial Z^{out}}{\partial D_{out}} \cdot \frac{\partial D_{out}}{\partial FL_{d2}} = \left[\sigma(Z^{out}_i) - y_i\right] \cdot W^o \cdot \left[(1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma^2_{(X)} + \epsilon}} - \frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma^2_{(X)} + \epsilon)^{3/2}}\right] \\
\frac{\partial \text{Loss}}{\partial D_{t2}} &= \frac{\partial \text{Loss}}{\partial \sigma(Z^{out})} \cdot \frac{\partial \sigma(\partial Z^{out})}{\partial Z^{out}} \cdot \frac{\partial Z^{out}}{\partial D_{out}} \cdot \frac{\partial D_{out}}{\partial D_{t2}} = \left[(\sigma(Z^{out}_i) - y_i\right] \cdot W^o \cdot \left[(1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma^2_{(X)} + \epsilon}} - \frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma^2_{(X)} + \epsilon)^{3/2}}\right]
\end{aligned}$$