```
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_{fl_1}^e + b_{fl_1}^e)
                                                      FL_{e2} = FL_{e1} * W_{fl_2}^e + b_{fl_2}^e
                                                      Ec_{out} = Norm(FL_{e2} + Ec_{t1})
                                              Input^d = ([Input_{ohe}]_{n_W \times n_V}^d \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_t}}])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_t}} \right) V_C + D_{t1} \right) = Norm (A_{cr} + D_{t1})
                                                      FL_{d1} = ReLU(D_{t2} \cdot W_{fl_1}^d + b_{fl_1}^d)
                                                    FL_{d2} = FL_{d1} \cdot W_{fl_2}^d + b_{fl_2}^d
                                                     D_{out} = Norm(FL_{d2} + D_{t2}) = \frac{(FL_{d2} + D_{t2}) - \mu_{(FL_{d2} + D_{t2})}}{\sqrt{\sigma_{(FL_{d2} + D_{t2})}^2 + \epsilon}}
                                                          Z_{out} = D_{out} \cdot W^o + b^o
                                            \sigma(Z^{out}) = \sigma(D_{out} \cdot W^o + b^o)
                                                 \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_i^{out}) - 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_i^{out}) - y_i) \cdot W^o
                                                 \frac{\partial \text{Loss}}{\partial W^0} = \frac{\partial \text{Loss}}{\partial Z^{out}} \cdot \frac{\partial Z^{out}}{\partial W^0} = (\sigma(Z_i^{out}) - y_i) \cdot D_{out}
                                                 \frac{\partial \text{Loss}}{\partial b^0} = \frac{\partial \text{Loss}}{\partial Z^{out}} \cdot \frac{\partial Z^{out}}{\partial b^0} = (\sigma(Z_i^{out}) - y_i) \cdot 1
                                                \frac{\partial \text{Loss}}{\partial z} = \frac{\partial \text{Loss}}{\partial z} \cdot \frac{\partial \sigma(\partial Z^{out})}{\partial z} \cdot \frac{\partial Z^{out}}{\partial z} \cdot \frac{\partial D_{out}}{\partial z}
                                               \frac{\partial FL_{d2}}{\partial FL_{d2}} = \frac{\partial \sigma(Z^{out})}{\partial \sigma(Z^{out})} \cdot \frac{\partial Z^{out}}{\partial Z^{out}} \cdot \frac{\partial Z^{out}}{\partial D_{out}} \cdot \frac{\partial Z^{out}}{\partial FL_{d2}}
                                                 \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}}
                                              \frac{\partial D_{out}}{\partial F L_{d2}} = \frac{\partial Norm(F L_{d2} + D_{t2})}{\partial F L_{d2}} = \frac{\partial \frac{(F L_{d2} + D_{t2}) - \mu_{(F L_{d2} + D_{t2})}}{\sqrt{\sigma_{(F L_{d2} + D_{t2})}^2 + \epsilon}}}{\partial F L_{d2}}
                                                                   X = FL_{d2} + D_{t2};
                                                 \frac{\partial D_{out}}{\partial D_{t2}} = \frac{\sqrt{\sigma_{(X)}^2 + \epsilon}}{\partial X} \cdot \frac{\partial X}{\partial D_{t2}}
                                  \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma_{(X)}^2 + \epsilon}}}{\partial X} = \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma_{(X)}^2 + \epsilon}}}{\partial X} + \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma_{(X)}^2 + \epsilon}}}{\partial \mu_{(X)}} \cdot \frac{\partial \mu_{(X)}}{\partial X} + \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma_{(X)}^2 + \epsilon}}}{\partial \sigma_{(X)}^2} \cdot \frac{\partial \sigma_{(X)}^2}{\partial X}
                                \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma_{(X)}^2 + \epsilon}}}{\partial X} = \frac{1}{\sqrt{\sigma_{(X)}^2 + \epsilon}}
\frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma_{(X)}^2 + \epsilon}}}{\partial \mu_{(X)}} \cdot \frac{\partial \mu_{(X)}}{\partial X} = -\frac{1}{\sqrt{\sigma_{(X)}^2 + \epsilon}} \cdot \frac{1}{N}
                                                        \sigma_{(X)}^2 = \frac{1}{N} \sum (X - \mu_x)^2; \quad \frac{\partial \sigma_{(X)}^2}{\partial X} = \frac{2}{N} (X - \mu_x) \cdot 1
 \frac{\partial \frac{X - \mu_{(X)}}{\sqrt{\sigma_{(X)}^2 + \epsilon}}}{\partial \sigma_{(Y)}^2} \cdot \frac{\partial \sigma_{(X)}^2}{\partial X} = \frac{\partial}{\partial \sigma_{(Y)}^2} \frac{X - \mu_{(X)}}{(\sigma_{(Y)}^2 + \epsilon)^{-1/2}} \cdot \frac{\partial \sigma_{(X)}^2}{\partial X} = -\frac{1}{2} (\sigma_{(X)}^2 + \epsilon)^{-3/2} \cdot 1 \cdot (X - \mu_{(X)}) \frac{\partial \sigma_{(X)}^2}{\partial X} = -\frac{1}{2} \frac{X - \mu_{(X)}}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_x)^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_x)^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_x)^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x) \cdot 1 = -\frac{1}{N} \frac{(X - \mu_x)^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \cdot \frac{2}{N} (X - \mu_x)^2 \cdot
                                             \frac{\partial D_{out}}{\partial F L_{d2}} = (1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma_{(X)}^2 + \epsilon}} - \frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} = \frac{\partial D_{out}}{\partial D_{t2}}
                                              \frac{\partial \text{Loss}}{\partial F L_{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 F L_{d2}} = \left[\sigma(Z_i^{out}) - y_i\right] \cdot W^o \cdot \left[(1 - \frac{1}{N})\frac{1}{\sqrt{\sigma_{(X)}^2 + \epsilon}} - \frac{1}{N}\frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \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_i^{out}) - y_i) \cdot W^o \cdot \left[ (1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma_{(X)}^2 + \epsilon}} - \frac{1}{N} \frac{(X - \mu_{(X)})^2}{(\sigma_{(X)}^2 + \epsilon)^{3/2}} \right] \right]
                                              \frac{\partial \text{Loss}}{\partial F L_{d1}} = \frac{\partial \text{Loss}}{\partial F L_{d2}} \cdot \frac{\partial F L_{d2}}{\partial F L_{d1}} = \frac{\partial \text{Loss}}{\partial F L_{d2}} \cdot W_{fl2}^d

\frac{\partial \text{Loss}}{\partial W_{fl2}^d} = \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot \frac{\partial FL_{d2}}{\partial W_{fl2}^d} = \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot FL_{d1}

\frac{\partial \text{Loss}}{\partial b_{fl2}^d} = \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot \frac{\partial FL_{d2}}{\partial b_{fl2}^d} = \frac{\partial \text{Loss}}{\partial FL_{d2}} \cdot 1

                                                        X_{1d} = D_{t2} \cdot W_{fl_1}^d + b_{fl_1}^d
                                                 \frac{\partial \text{Loss}}{\partial D_{t2}} = \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot \frac{\partial FL_{d1}}{\partial ReLu(X_{1d})} \cdot \frac{\partial ReLu(X_{1d})}{\partial X_{1d}} \cdot \frac{\partial X_{1d}}{\partial D_{t2}} = \begin{cases} \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot W_{fl_1}^d, & X_{1d} > 0\\ 0, & \text{otherwise} \end{cases}
                                                \frac{\partial \text{Loss}}{\partial W_{fl1}} = \frac{\partial \text{Loss}}{\partial F L_{d1}} \cdot \frac{\partial F L_{d1}}{\partial Re Lu(X_{1d})} \cdot \frac{\partial Re Lu(X_{1d})}{\partial X_{1d}} \cdot \frac{\partial X_{1d}}{\partial W_{fl1}} = \begin{cases} \frac{\partial \text{Loss}}{\partial F L_{d1}} \cdot D_{t2}, & X_{1d} > 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(X_{1d})} \cdot \frac{\partial ReLu(X_{1d})}{\partial X_{1d}} \cdot \frac{\partial X_{1d}}{\partial b_{fl1}} = \begin{cases} \frac{\partial \text{Loss}}{\partial FL_{d1}} \cdot 1, & X_{1d} > 0\\ 0, & \text{otherwise} \end{cases}
                                                          A_{cr} = \sigma \left( \frac{Q_C K_C^T}{\sqrt{d_L}} \right) \cdot V_C
                                                          Res = A_{cr} + D_{t1}
                                                 \frac{\partial \text{Loss}}{\partial A_{cr}} = \frac{\partial \text{Loss}}{\partial D_{t2}} \cdot \frac{\partial D_{t2}}{\partial Norm(Res)} \cdot \frac{\partial Norm(Res)}{\partial Res} \cdot \frac{\partial Res}{\partial A_{cr}} = \frac{\partial \text{Loss}}{\partial D_{t2}} \cdot \left[ (1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma_{(Res)}^2 + \epsilon}} - \frac{1}{N} \frac{(Res - \mu_{(Res)})^2}{(\sigma_{(Res)}^2 + \epsilon)^{3/2}} \right]
                                                  \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_{t1}^{qc}} = \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 E c_{cout}^K} = \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \frac{\partial A_{cr}}{\partial K_C} \cdot \frac{\partial K_C}{\partial E c_{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 E c_{cout}^{V}} = \frac{\partial \text{Loss}}{\partial A_{cr}} \cdot \frac{\partial A_{cr}}{\partial V_{C}} \cdot \frac{\partial V_{C}}{\partial E c_{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 E c_{cout}} = \frac{\partial \text{Loss}}{\partial E c_{cout}^K} + \frac{\partial \text{Loss}}{\partial E c_{cout}^V}
                                                A_{mask} = \sigma(Mask[\frac{Q_D K_D^T}{\sqrt{d_L}}])V_D = \sigma(\frac{Q_D K_D^T}{\sqrt{d_L}} + Mask)V_D
                                                               X_d = Input^d + A_{masked}
                                             \frac{\partial \text{Loss}}{\partial D_{t1}} = \frac{\partial \text{Loss}}{\partial D_{t1}} \cdot \frac{\partial Norm(X_d)}{\partial D_{t1}}
                                             \frac{\partial A_{mask}}{\partial A_{mask}} = \frac{\partial D_{t1}}{\partial D_{t1}} \cdot \frac{\partial Norm(X_d)}{\partial Norm(X_d)} \cdot \frac{\partial A_{mask}}{\partial A_{mask}}
                                                                                                                                                                                          \partial Norm(X_d)
                                                                                        \partial \text{Loss} \qquad \partial D_{t1}
                                  \frac{1}{\partial input_{dt1}^d} = \frac{1}{\partial D_{t1}} \cdot \frac{1}{\partial Norm(X_d)} \cdot \frac{1}{\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_d^q} = \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_d^k} = \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_d^v} = \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^d} = \frac{\partial \text{Loss}}{\partial input_d^q} + \frac{\partial \text{Loss}}{\partial input_d^k} + \frac{\partial \text{Loss}}{\partial input_d^v} + \frac{\partial \text{Loss}}{\partial input_{dt1}^d}
                                             \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
                                           \frac{\partial \text{Loss}}{\partial E c_{cout}} = \frac{\partial \text{Loss}}{\partial E c_{cout}^K} + \frac{\partial \text{Loss}}{\partial E c_{cout}^V}
                                                          X_{e2} = FL_{e2} + Ec_{t1}
                                              \frac{\partial \text{Loss}}{\partial F L_{e2}} = \frac{\partial \text{Loss}}{\partial E c_{cout}} \cdot \frac{\partial E c_{cout}}{\partial Norm(X_{e2})} \cdot \frac{\partial Norm(X_{e2})}{\partial F L_{e2}} = \frac{\partial \text{Loss}}{\partial E c_{cout}} \cdot \left[ (1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma_{(X_{e2})}^2 + \epsilon}} - \frac{1}{N} \frac{(X_{e2} - \mu_{(X_{e2})})^2}{(\sigma_{(X_{e2})}^2 + \epsilon)^{3/2}} \right]
                                             \frac{\partial \text{Loss}}{\partial E^n c_{t1}} = \frac{\partial \text{Loss}}{\partial E c_{cout}} \cdot \frac{\partial E c_{cout}}{\partial Norm(X_{e2})} \cdot \frac{\partial Norm(X_{e2})}{\partial E c_{t1}} = \frac{\partial \text{Loss}}{\partial E c_{cout}} \cdot \left[ (1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma_{(X_{e2})}^2 + \epsilon}} - \frac{1}{N} \frac{(X_{e2} - \mu_{(X_{e2})})^2}{(\sigma_{(X_{e2})}^2 + \epsilon)^{3/2}} \right]

\frac{\partial \text{Loss}}{\partial W_{fl2}^e} = \frac{\partial \text{Loss}}{\partial F L_{e2}} \cdot \frac{\partial F L_{e2}}{\partial W_{fl2}^e} = \frac{\partial \text{Loss}}{\partial F L_{e2}} \cdot F l_{e1}

\frac{\partial \text{Loss}}{\partial b_{fl2}^e} = \frac{\partial \text{Loss}}{\partial F L_{e2}} \cdot \frac{\partial F L_{e2}}{\partial b_{fl2}^e} = \frac{\partial \text{Loss}}{\partial F L_{e2}} \cdot 1

                                              \frac{\partial \text{Loss}}{\partial F L_{e1}} = \frac{\partial \text{Loss}}{\partial F L_{e2}} \cdot \frac{\partial F L_{e2}}{\partial F L_{e1}} = \frac{\partial \text{Loss}}{\partial F l_{e2}} \cdot W_{fl2}^e
                                                          X_{e1} = Ec_{t1} \cdot W_{fl_1}^e + b_{fl_1}^e
                                                 \frac{\partial \text{Loss}}{\partial E_{ct1}} = \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot \frac{\partial FL_{e1}}{\partial ReLu(X_{e1})} \cdot \frac{\partial ReLu(X_{e1})}{\partial X_{e1}} \cdot \frac{\partial X_{e1}}{\partial Ec_{t1}} = \begin{cases} \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot W_{fl_1}^e, & X_{e1} > 0\\ 0, & \text{otherwis} \end{cases}
                                               \frac{\partial \text{Loss}}{\partial W_{fl1}^e} = \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot \frac{\partial FL_{e1}}{\partial ReLu(X_{e1})} \cdot \frac{\partial ReLu(X_{e1})}{\partial (X_{e1})} \cdot \frac{\partial (X_{e1})}{\partial W_{fl1}^e} = \begin{cases} \frac{\partial \text{Loss}}{\partial FL_{e1}} \cdot Ec_{t1}, & X_{e1} > 0\\ 0, & \text{otherwise} \end{cases}
                                                 \frac{\partial \text{Loss}}{\partial b_{fl1}^e} = \frac{\partial \text{Loss}}{\partial F L_{e1}} \cdot \frac{\partial F L_{e1}}{\partial ReLu(X_{e1})} \cdot \frac{\partial ReLu(X_{e1})}{\partial (X_{e1})} \cdot \frac{\partial (X_{e1})}{\partial b_{fl1}^e} = \begin{cases} \frac{\partial \text{Loss}}{\partial F L_{e1}} \cdot 1, & X_{e1} > 0\\ 0, & \text{otherwise} \end{cases}
                                                             A_e = \sigma(\frac{Q_E K_E^T}{\sqrt{d_{th}}}) V_E
                                                            X_{e1} = A_e + Input^e
                                                \frac{\partial \text{Loss}}{\partial A_e} = \frac{\partial \text{Loss}}{\partial E c_{t1}} \cdot \frac{\partial E c_{t1}}{\partial Norm(X_{e1})} \cdot \frac{\partial Norm(X_{e1})}{\partial A_e} = \frac{\partial \text{Loss}}{\partial E_{ct1}} \cdot \left[ (1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma_{(X_{e1})}^2 + \epsilon}} - \frac{1}{N} \frac{(X_{e1} - \mu_{(X_{e1})})^2}{(\sigma_{(X_{e1})}^2 + \epsilon)^{3/2}} \right]
                                     \frac{\partial \text{Loss}}{\partial Input_n^e} = \frac{\partial \text{Loss}}{\partial Ec_{t1}} \cdot \frac{\partial Ec_{t1}}{\partial Norm(X_{e1})} \cdot \frac{\partial Norm(X_{e1})}{\partial Input^e} = \frac{\partial \text{Loss}}{\partial E_{ct1}} \cdot \left[ (1 - \frac{1}{N}) \frac{1}{\sqrt{\sigma_{(X_{e1})}^2 + \epsilon}} - \frac{1}{N} \frac{(X_{e1} - \mu_{(X_{e1})})^2}{(\sigma_{(X_{e1})}^2 + \epsilon)^{3/2}} \right]
                                                 \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_Q^e} = \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_K^e} = \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_{V}^{e}} = \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
```

 $Input^e = ([Input_{ohe}]_{n_W \times n_V}^e \bullet [W_{emb}]_{n_V \times n_E}) + P_e$