D3PM Loss Function

The t-th diffusion loss term:

$$
\begin{aligned}
& L(\boldsymbol{x}_0, t) \\
= \;& \mathbb{E}_{q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}[D_{\mathrm{KL}}[q(\boldsymbol{x}_t|\,\boldsymbol{x}_{t+1}, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_t|\,\boldsymbol{x}_{t+1})]] \\
= \;& \mathbb{E}_{q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}[D_{\mathrm{KL}}[q(\boldsymbol{x}_t|\,\boldsymbol{x}_{t+1}, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_t|\,\boldsymbol{x}_{t+1})]] \\
= \;& \mathbb{E}_{q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left[\prod_{c=1}^{C} q(\boldsymbol{x}_t^c|\,\boldsymbol{x}_{t+1}^c, \boldsymbol{x}_0^c)\middle|\middle|\prod_{k=1}^{K} p_\theta(\boldsymbol{x}_t^c|\,\boldsymbol{x}_{t+1})\right]\right] \\
= \;& \mathbb{E}_{q(\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[\sum_{c=1}^{C} D_{\mathrm{KL}}[q(\boldsymbol{x}_t^c|\,\boldsymbol{x}_{t+1}^c, \boldsymbol{x}_0^c)||p_\theta(\boldsymbol{x}_t^c|\,\boldsymbol{x}_{t+1})]\right]
\end{aligned}
$$

Given $\boldsymbol{x}_{t+1}$ and $\boldsymbol{x}_0$, $\{q(\boldsymbol{x}_t^c|\,\boldsymbol{x}_{t+1}^c, \boldsymbol{x}_0^c)\}$ and $\{p_\theta(\boldsymbol{x}_t^c|\,\boldsymbol{x}_{t+1})\}$ can each be represented by a matrix of size $(C, \#\text{tokens})$.