> **Q3** : The equation in 2.2 misses a “+” after const.? Also, how do we get the equation 4?

A3: Thank you for pointing it out! We rewrite the equation 2.2.
For equation 4, we follow [1] and derive this equation as below:

[1] introduces the reparameterized trick to discrete diffusion model, and the backward transition $q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)})$ can be rewritten as:

$$
\begin{aligned}
&q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)}) \\
&= \begin{cases} \lambda_{t-1}^{(1)}\boldsymbol{x}^{(t)} + (1-\lambda_{t-1}^{(1)})\boldsymbol{q}_{\text{noise}}, & \text{if } \boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t)} \\ \lambda_{t-1}^{(2)}\boldsymbol{x}^{(t)} + (1-\lambda_{t-1}^{(2)})\boldsymbol{q}_{\text{noise}}(\boldsymbol{x}^{(t)}), & \text{if } \boldsymbol{x}^{(t)} \neq \boldsymbol{x}^{(0)} \end{cases}
\end{aligned}
$$

where $\boldsymbol{q}_{\text{noise}}(\boldsymbol{x}^{(t)}) = \beta_t\boldsymbol{x}^{(t)} + (1-\beta_t)\boldsymbol{q}_{\text{noise}}$, and both $\lambda_{t-1}^{(1)}$ and $\lambda_{t-1}^{(2)}$ are constants relating to β_t and β_{t-1}.
Sampling from it is equivalent to first sampling from a Bernoulli distribution and then the corresponding component distribution:

$$
\begin{aligned}
v_{t-1}^{(1)} &\sim \text{Bernoulli}\left(\lambda_{t-1}^{(1)}\right), & \boldsymbol{u}_t^{(1)} &\sim \texttt{Cat}\left(\boldsymbol{u}; \boldsymbol{p} = \boldsymbol{q}_{\text{noise}}\right) \\
v_{t-1}^{(2)} &\sim \text{Bernoulli}\left(\lambda_{t-1}^{(2)}\right), & \boldsymbol{u}_t^{(2)} &\sim \texttt{Cat}\left(\boldsymbol{u}; \boldsymbol{p} = \boldsymbol{q}_{\text{noise}}(\boldsymbol{x}_t)\right)
\end{aligned}
$$

$$
\boldsymbol{x}_{t-1} = \begin{cases} v_{t-1}^{(1)}\boldsymbol{x}_t + \left(1 - v_{t-1}^{(1)}\right)\boldsymbol{u}_t^{(1)}, & \text{if } \boldsymbol{x}_t = \boldsymbol{x}_0 \\ v_{t-1}^{(2)}\boldsymbol{x}_t + \left(1 - v_{t-1}^{(2)}\right)\boldsymbol{u}_t^{(2)}, & \text{if } \boldsymbol{x}_t \neq \boldsymbol{x}_0 \end{cases}
$$

This reparameterizes the backward transitions $q(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)})$ and $p_\theta(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)})$ into $q(\boldsymbol{x}^{(t-1)}, \boldsymbol{v}^{(t-1)}|\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(0)})$ and $p_\theta(\boldsymbol{x}^{(t-1)}, \boldsymbol{v}^{(t-1)}|\boldsymbol{x}^{(t)})$, respectively.

Since each token is modeled **conditionally independently**, so we consider the backward transition for **each token**, and sum the losses for them.
For i-th position, the backward transition is $q(\boldsymbol{x}_i^{(t-1)}, \boldsymbol{v}_i^{(t-1)}|\boldsymbol{x}_i^{(t)}, \boldsymbol{x}_i^{(0)})$.

As shown in [1] (appendix C), the loss at i-th token can be written as below:

$$
\mathcal{J}_{t,i} = \mathbf{E}_{q(\boldsymbol{v}_i^{(t-1)})}\left[KL[q(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{v}_i^{(t-1)}, \boldsymbol{x}_i^{(t)}, \boldsymbol{x}_i^{(0)})||p_\theta(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{v}_i^{(t-1)}, \boldsymbol{x}_i^{(t)})]\right]
$$

Let $b_i(t) = \mathbf{1}_{x_i^{(t)} = x_i^{(0)}}$, $q(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{v}_i^{(t-1)}, \boldsymbol{x}_i^{(t)}, \boldsymbol{x}_i^{(0)})$ can be written as:

$$
\begin{aligned}
&q(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{v}_i^{(t-1)}, \boldsymbol{x}_i^{(t)}, \boldsymbol{x}_i^{(0)}) \\
&= \begin{cases} v_{t-1,i}^{(1)}\boldsymbol{x}_i^{(t)} + (1 - v_{t-1,i}^{(1)})\boldsymbol{q}_{\text{noise}} & \text{if } b_i(t) = 0, \\ v_{t-1,i}^{(2)}\boldsymbol{x}_i^{(0)} + (1 - v_{t-1,i}^{(2)})\boldsymbol{q}_{\text{noise}} & \text{if } b_i(t) = 1, \end{cases}
\end{aligned}
$$

And $p_\theta(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{v}_i^{(t-1)}, \boldsymbol{x}_i^{(t)})$ can be written as:

$$
\begin{aligned}
&p_\theta(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{v}_i^{(t-1)}, \boldsymbol{x}_i^{(t)}) \\
&= \begin{cases} v_{t-1,i}^{(1)}\boldsymbol{x}_i^{(t)} + (1 - v_{t-1,i}^{(1)})\boldsymbol{q}_{\text{noise}} & \text{if } b_i(t) = 0, \\ v_{t-1,i}^{(2)} p_\theta(\boldsymbol{x}_i^{(0)}|\boldsymbol{x}^{(t)}) + (1 - v_{t-1,i}^{(2)})\boldsymbol{q}_{\text{noise}} & \text{if } b_i(t) = 1, \end{cases}
\end{aligned}
$$

Therefore, the loss at i-th token can be computed by enumerating all cases with respect to $\boldsymbol{v}_i^{(t-1)}$ and $b_i(t)$. As noted in [1], the KL divergence is equal to $-\log p_\theta(x_i^{(0)}|x^{(t)})$ when $v_{t-1,i}^{(2)} = 1$ and $b_i(t) = 1$, while in other cases the KL divergence is 0.

So we have:

$$
\begin{aligned}
\mathcal{J}_t &= \sum_{1 \le i \le L} \mathcal{J}_{t,i} \\
&= \sum_{1 \le i \le L} \mathbf{E}_{q(\boldsymbol{v}_i^{(t-1)})} \left[KL[q(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{v}_i^{(t-1)}, \boldsymbol{x}_i^{(t)}, \boldsymbol{x}_i^{(0)})||p_\theta(\boldsymbol{x}_i^{(t-1)}|\boldsymbol{v}_i^{(t-1)}, \boldsymbol{x}_i^{(t)})] \right] \\
&= \sum_{1 \le i \le L} q(\boldsymbol{v}_i^{(t-1)} = 1)(-\log p_\theta(x_i^{(0)}|x^{(t)})) \\
&= -\lambda^{(t)} \sum_{1 \le i \le L} b_i(t) \cdot \log p_\theta(\boldsymbol{x}_i^{(0)}|\boldsymbol{x}^{(t)})
\end{aligned}
$$

[1] Zheng, L., Yuan, J., Yu, L., and Kong, L. A reparameterized discrete diffusion model for text generation. arXiv preprint arXiv:2302.05737