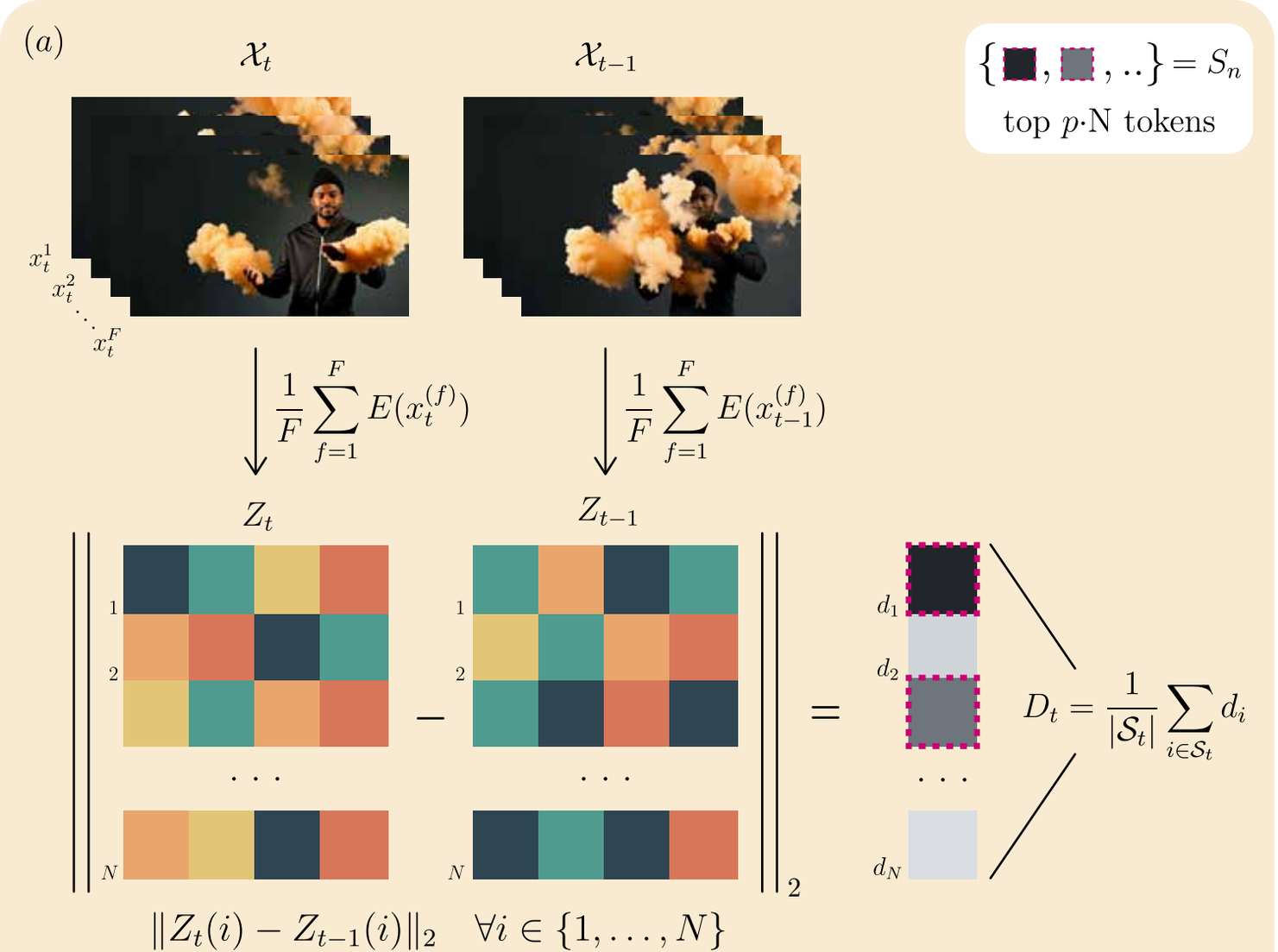
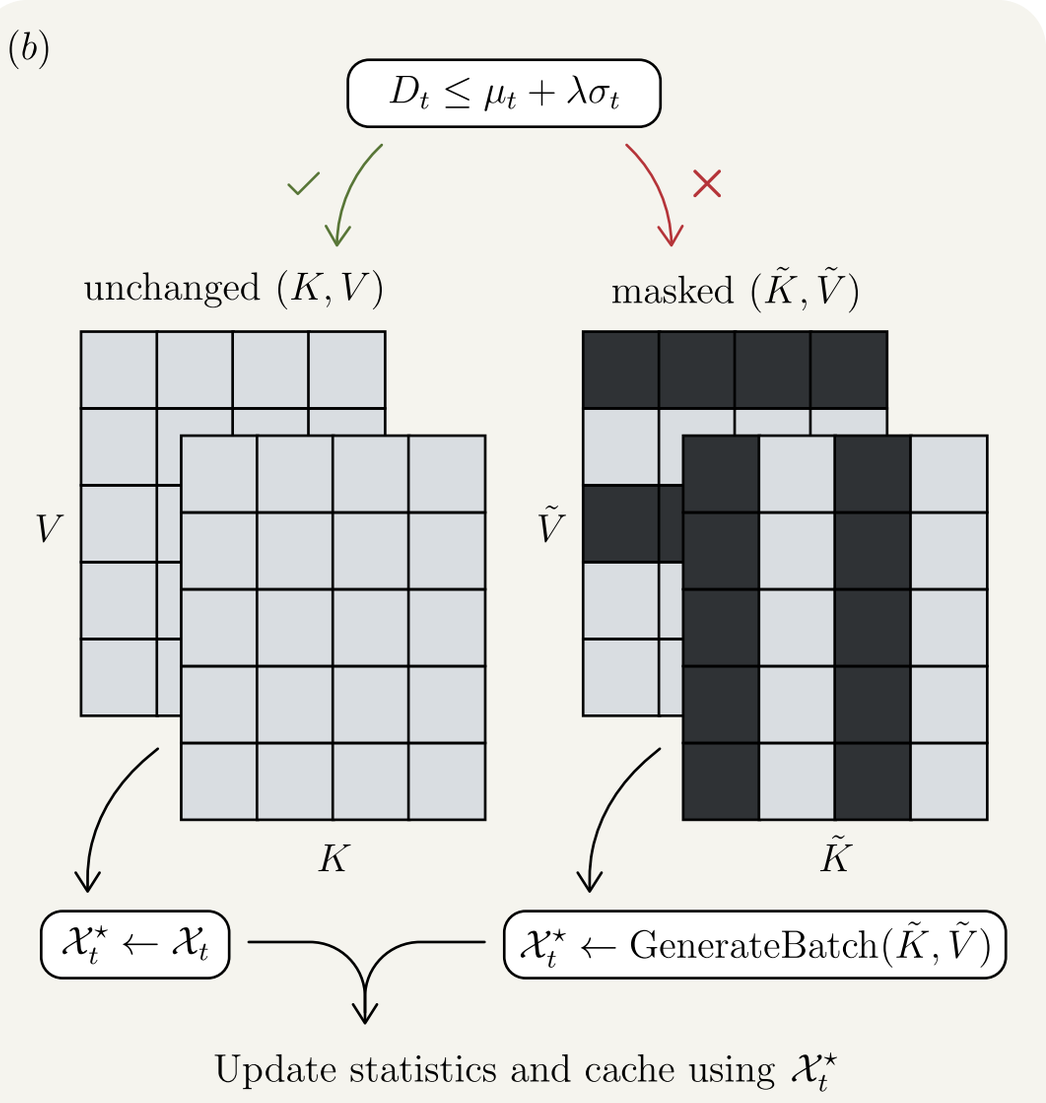

(a)
$\mathcal{X}_t$
$\mathcal{X}_{t-1}$
$\{\blacksquare, \blacksquare, ..\} = S_n$
top p·N tokens
x_t^1
x_t^2
x_t^F
$\frac{1}{F}\sum_{f=1}^{F} E(x_t^{(f)})$
$\frac{1}{F}\sum_{f=1}^{F} E(x_{t-1}^{(f)})$
Z_t
Z_{t-1}
1
2
N
$-$
$=$
d_1
d_2
d_N
$D_t = \frac{1}{|\mathcal{S}_t|}\sum_{i\in\mathcal{S}_t} d_i$
$\|Z_t(i) - Z_{t-1}(i)\|_2 \quad \forall i \in \{1, \ldots, N\}$


(b)
$D_t \le \mu_t + \lambda\sigma_t$
unchanged (K, V)
masked $(\tilde{K}, \tilde{V})$
V
$\tilde{V}$
K
$\tilde{K}$
$\mathcal{X}_t^\star \leftarrow \mathcal{X}_t$
$\mathcal{X}_t^\star \leftarrow \text{GenerateBatch}(\tilde{K}, \tilde{V})$
Update statistics and cache using $\mathcal{X}_t^\star$