(a) $\boldsymbol{W}^{KQ}, \sigma_w = 0.01, \sigma = 1$ (b) $\boldsymbol{W}^{PV}, \sigma_w = 0.01, \sigma = 1$ (c) Ratio of $\widehat{\boldsymbol{y}}_{T_{te}-1}/\boldsymbol{x}_{T_{te}}, \sigma = 1$

(d) $\boldsymbol{W}^{KQ}, \sigma_w = 0.01, c = 1$ (e) $\boldsymbol{W}^{PV}, \sigma_w = 0.01, c = 1$ (f) MSE $\|\widehat{\boldsymbol{y}}_{T_{te}-1} - \boldsymbol{x}_{T_{te}}\|_2^2, c = 1$

(g) $\boldsymbol{W}^{KQ}, a = b = 0.1, \sigma = 1$ (h) $\boldsymbol{W}^{PV}, a = b = 0.1, \sigma = 1$ (i) dynamics of $ab, \sigma = 1$

Figure 5: Figures in the first row are representative results of Gaussian start point with standard initialization. Figures in the second row are representative results of Example 4.1 with standard initialization. Figures in the last row are representative results of small-context scenarios ($d = T = 5$) with Gaussian start point and diagonal initialization, and the theoretical limit is $1/(5\sigma^2 + \frac{d-1}{T-2}\sigma^2 \sum_{t=2}^{T-1} \frac{1}{t-1})$ by Theorem 4.1. We note that we plot the read blocks in Assumption 3.1, which are related to the final prediction.