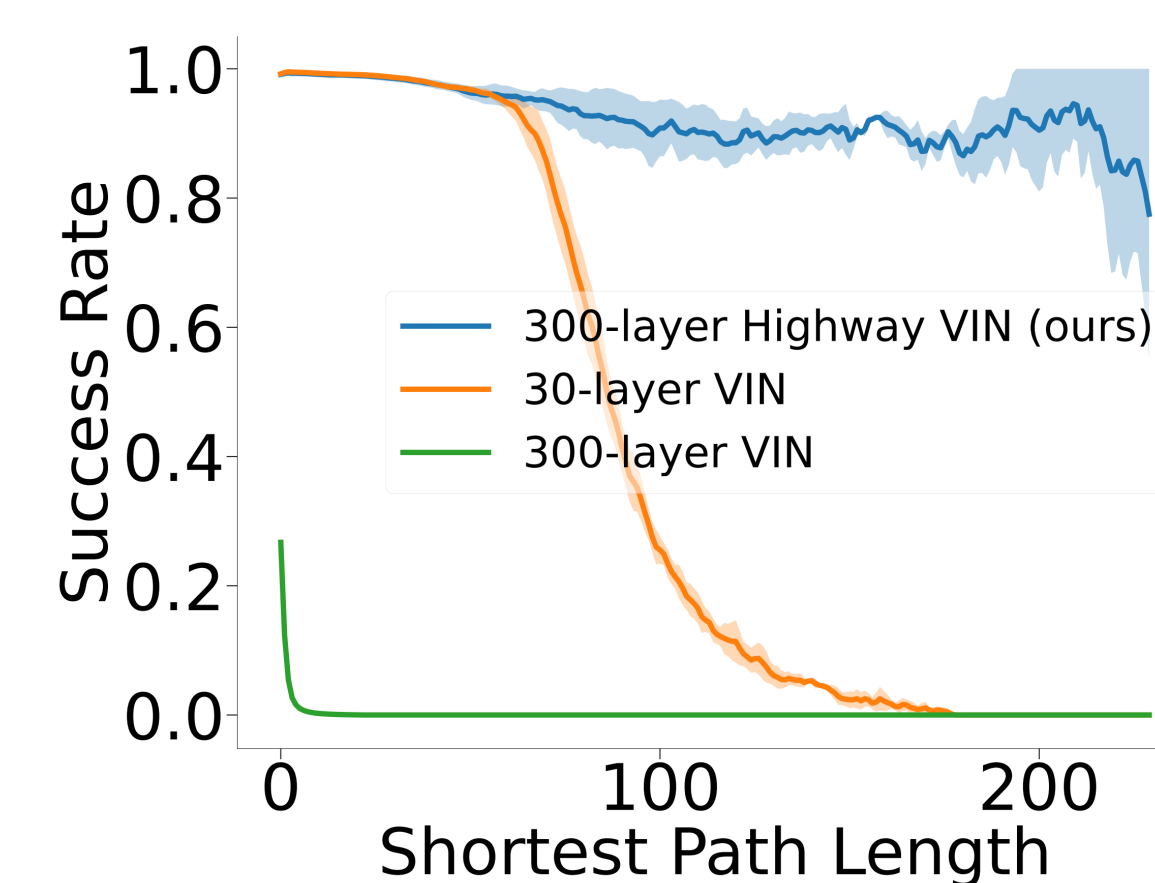
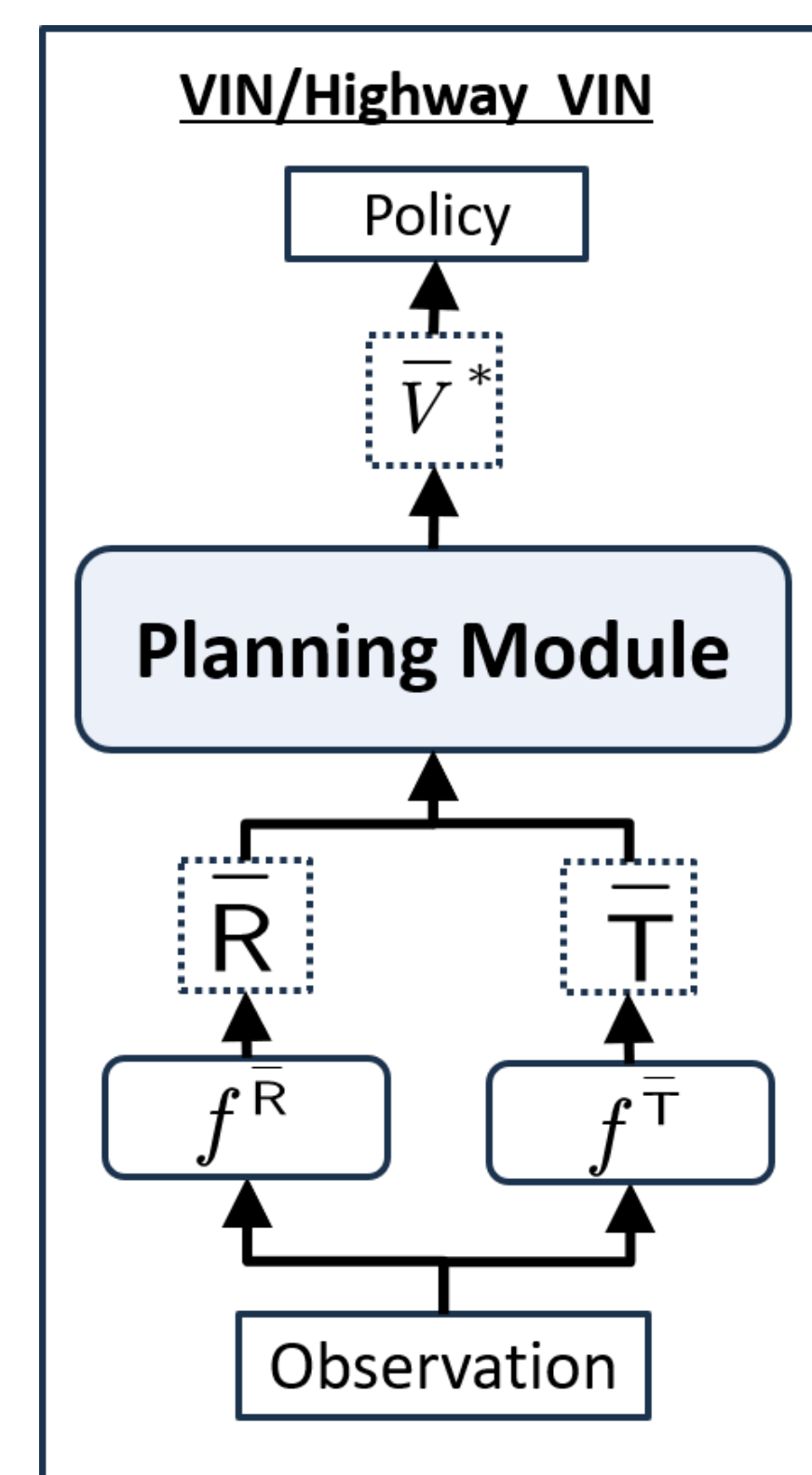
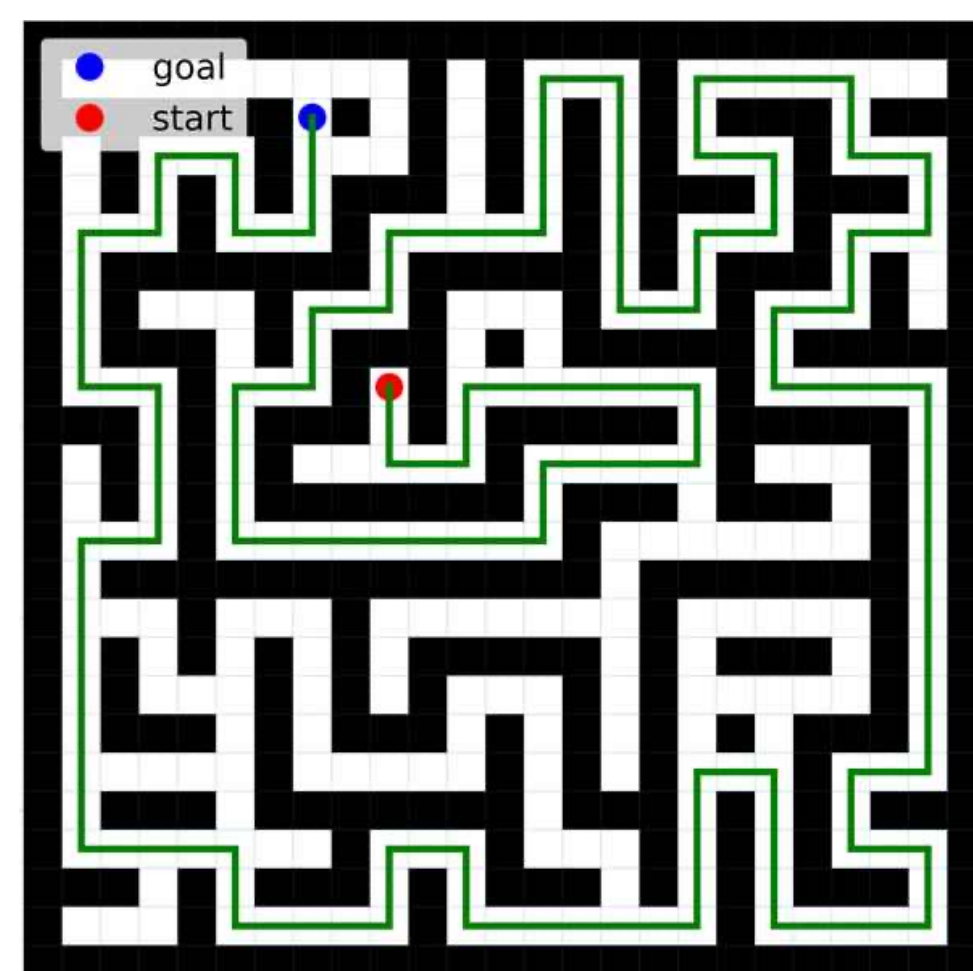
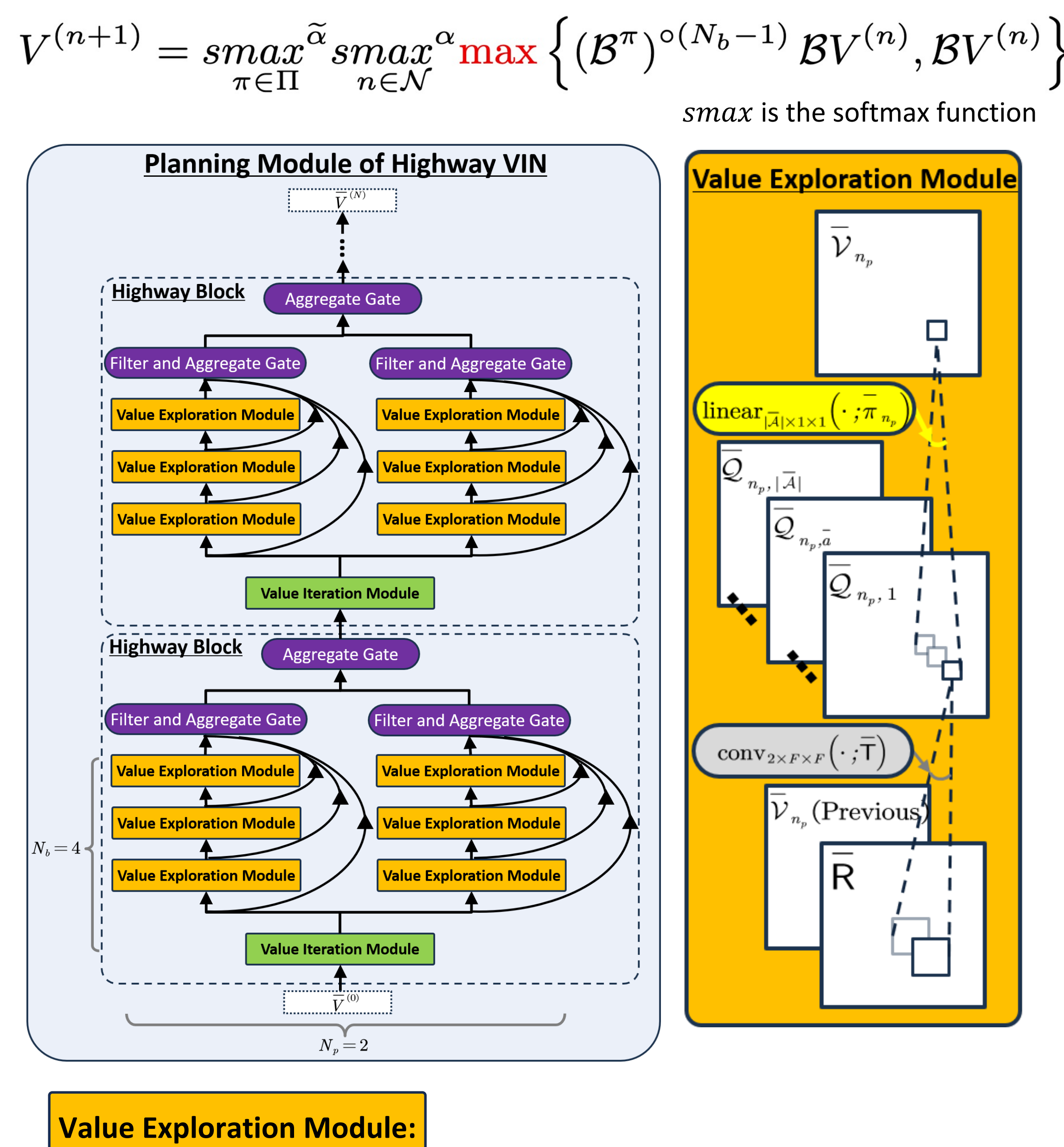
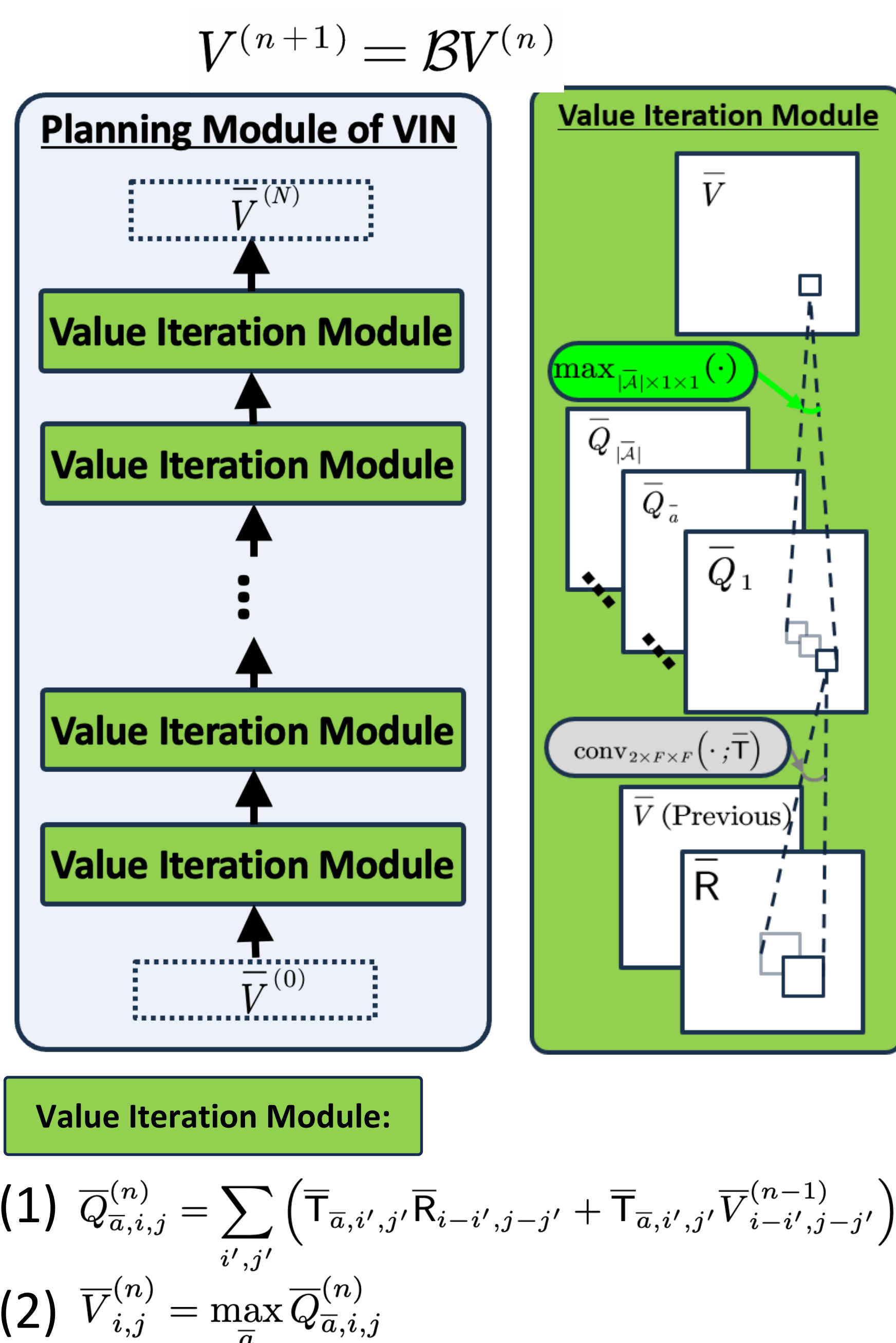


## Experiments



$$(BV)(s) \triangleq \max_a \sum_{s'} \mathcal{T}(s'|s, a) [\mathcal{R}(s, a, s') + \gamma V(s')], \quad (B^\pi V)(s) \triangleq \sum_a \pi(a|s) \sum_{s'} \mathcal{T}(s'|s, a) [\mathcal{R}(s, a, s') + \gamma V(s')]$$

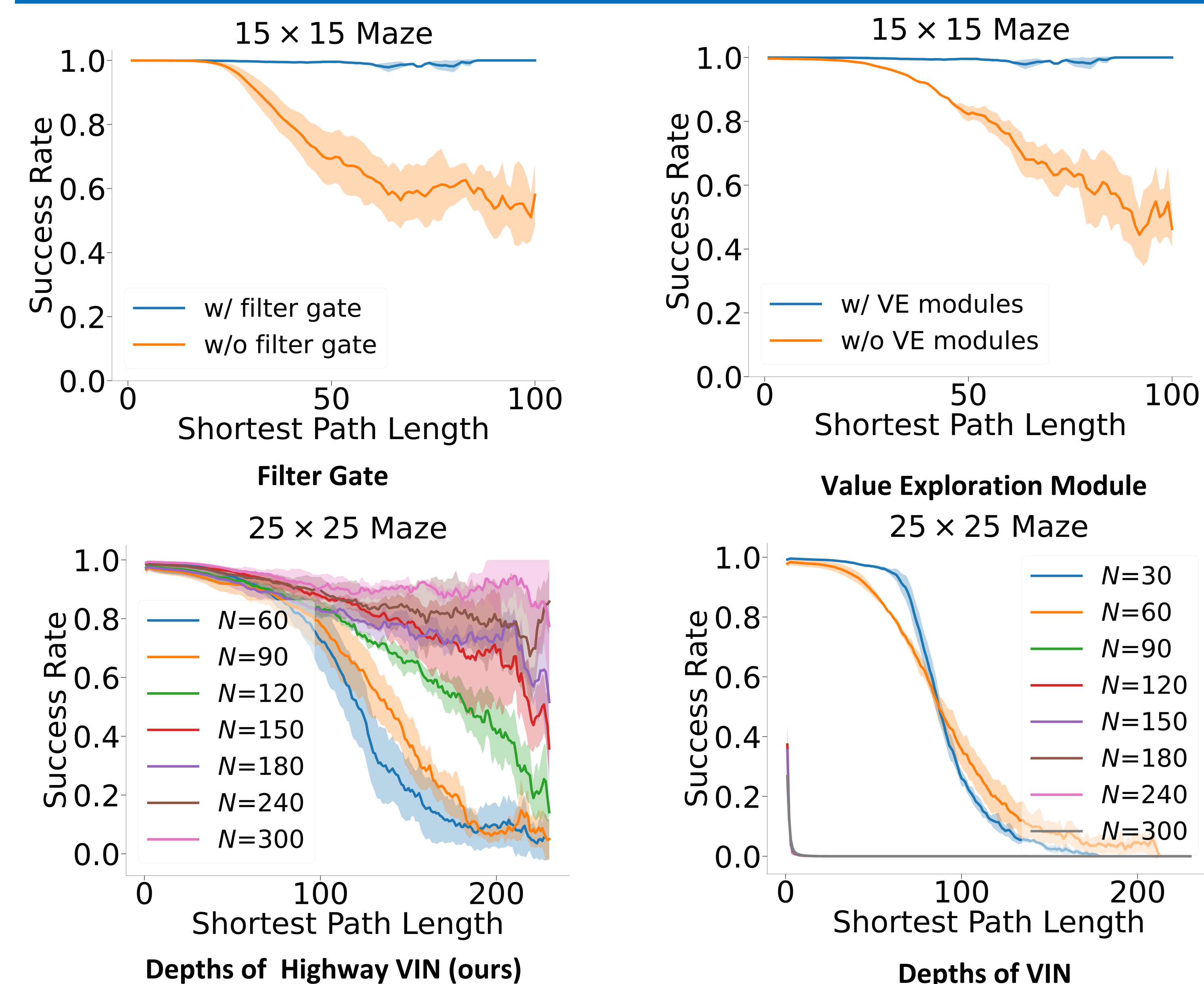
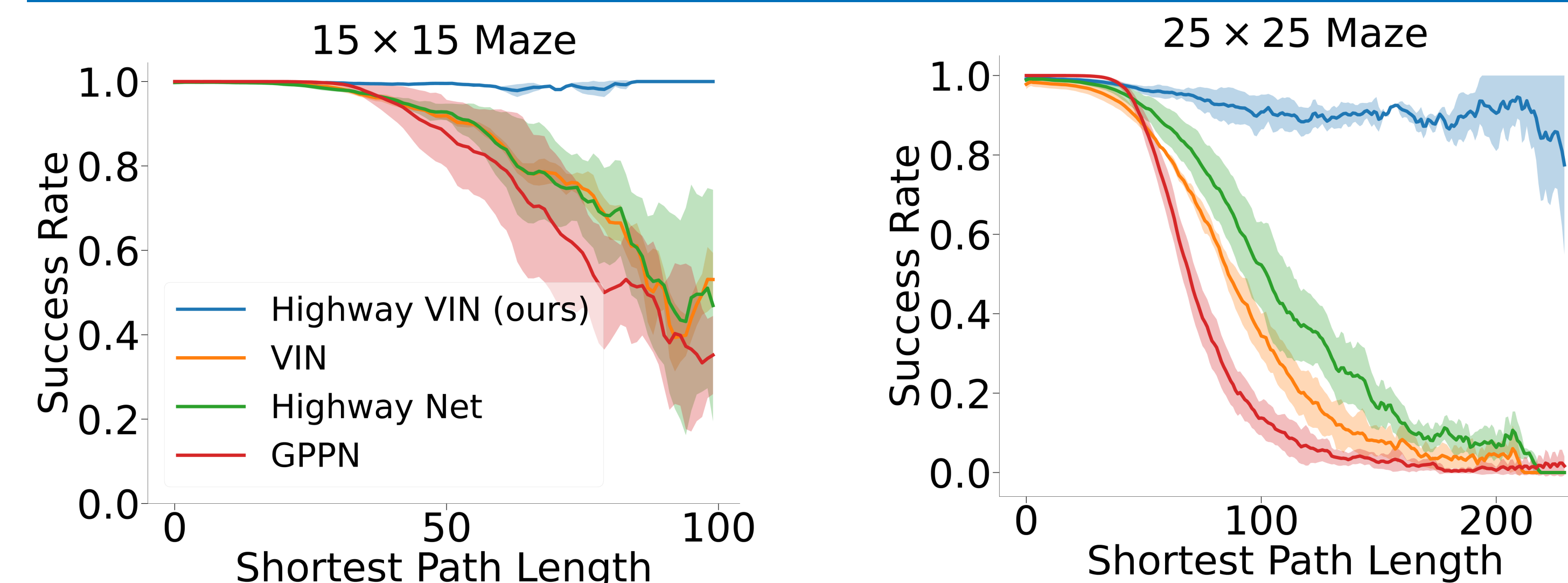


$$\begin{aligned} (1) \quad \overline{\mathcal{Q}}_{\overline{\pi}, \overline{a}, i, j}^{(n+n_b)} &= \sum_{i', j'} \left( \overline{\mathcal{T}}_{\overline{a}, i', j'} \overline{\mathcal{R}}_{-i', j-j'} + \overline{\mathcal{T}}_{\overline{a}, i', j'} \overline{\mathcal{V}}_{\overline{\pi}, i-i', j-j'}^{(n+n_b-1)} \right) \\ (2) \quad \overline{\mathcal{V}}_{n_p, \overline{a}, i, j}^{(n+n_b)} &= \sum_{\overline{a}} \overline{\pi}_{n_p, \overline{a}, i, j}^{(n+n_b)} \overline{\mathcal{Q}}_{n_p, \overline{a}, i, j}^{(n+n_b)} \end{aligned}$$

*embedded policy*

where  $\bar{\pi}_{n_p, \bar{a}, i, j}^{(n+n_b)} = \begin{cases} 1, & \bar{a} = \hat{a} \sim P\left(\cdot; \overline{\mathcal{Q}}_{n_p, \cdot, i, j}^{(n+n_b)}, \epsilon\right) \\ 0, & \text{otherwise,} \end{cases}$

$$P\left(\bar{a}; \overline{Q}_{n_p, \cdot, i, j}^{(n+n_b)}, \epsilon\right) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\bar{A}|}, & \bar{a} = \arg \max_{\bar{a}'} \overline{Q}_{n_p, \bar{a}', i, j}^{(n+n_b)} \\ \frac{\epsilon}{|\bar{A}|}, & \text{otherwise.} \end{cases}$$



Correspondence to: Yuhui Wang, yuhui.wang@kaust.edu.sa