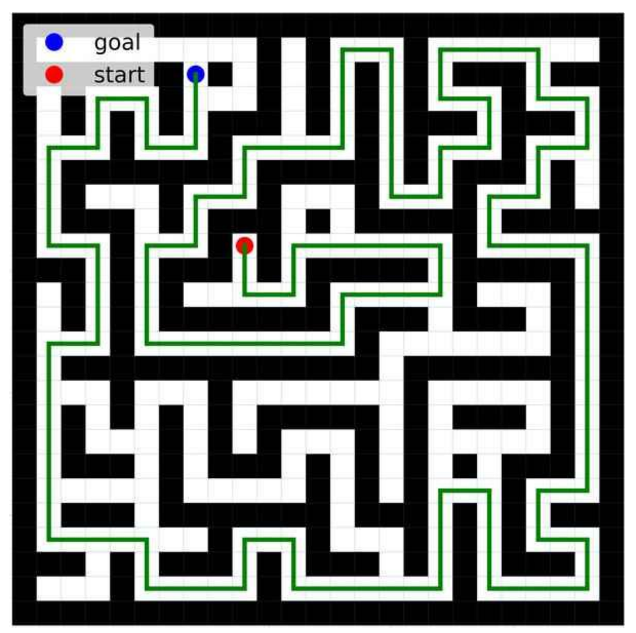
Highway Value Iteration Networks

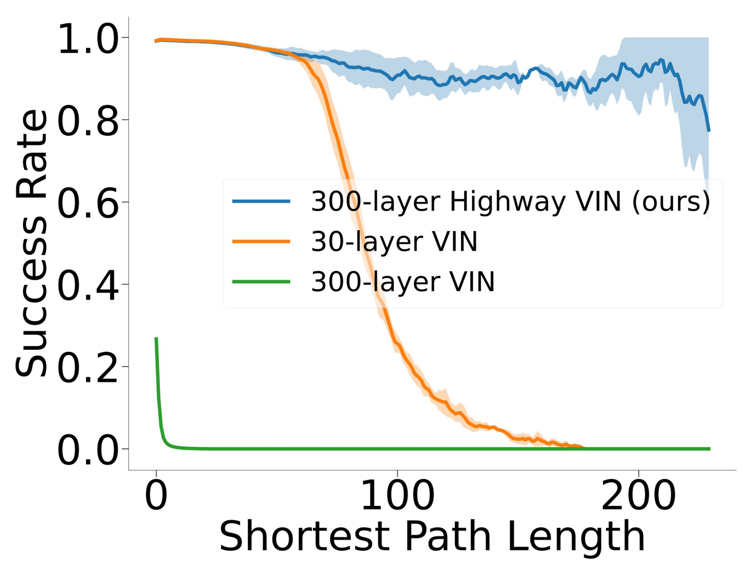
Yuhui Wang*[1], Weida Li*[2], Francesco Faccio[1,3], Qingyuan Wu[4], Jürgen Schmidhuber[1,3]

[1]AI Initiative, King Abudullah University of Science and Technology (KAUST), [2]National University of Singapore, [3]The Swiss AI Lab IDSIA/USI/SUPSI, [4]The University of Liverpool

Motivation

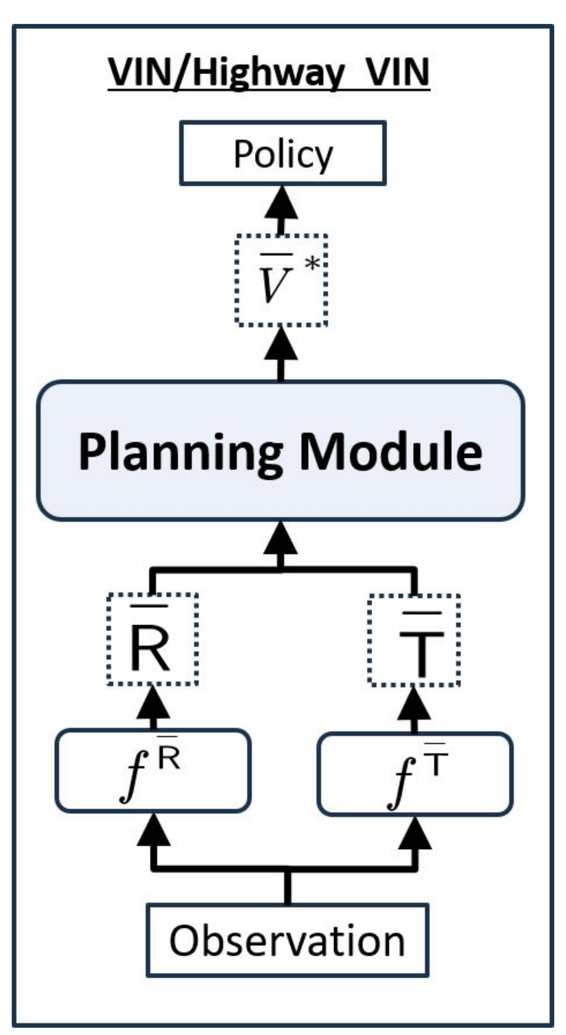


Maze Navigation



Performance on tasks with various shortest path lengths

Background



Architecture of Value Iteration Network (VIN) [1]

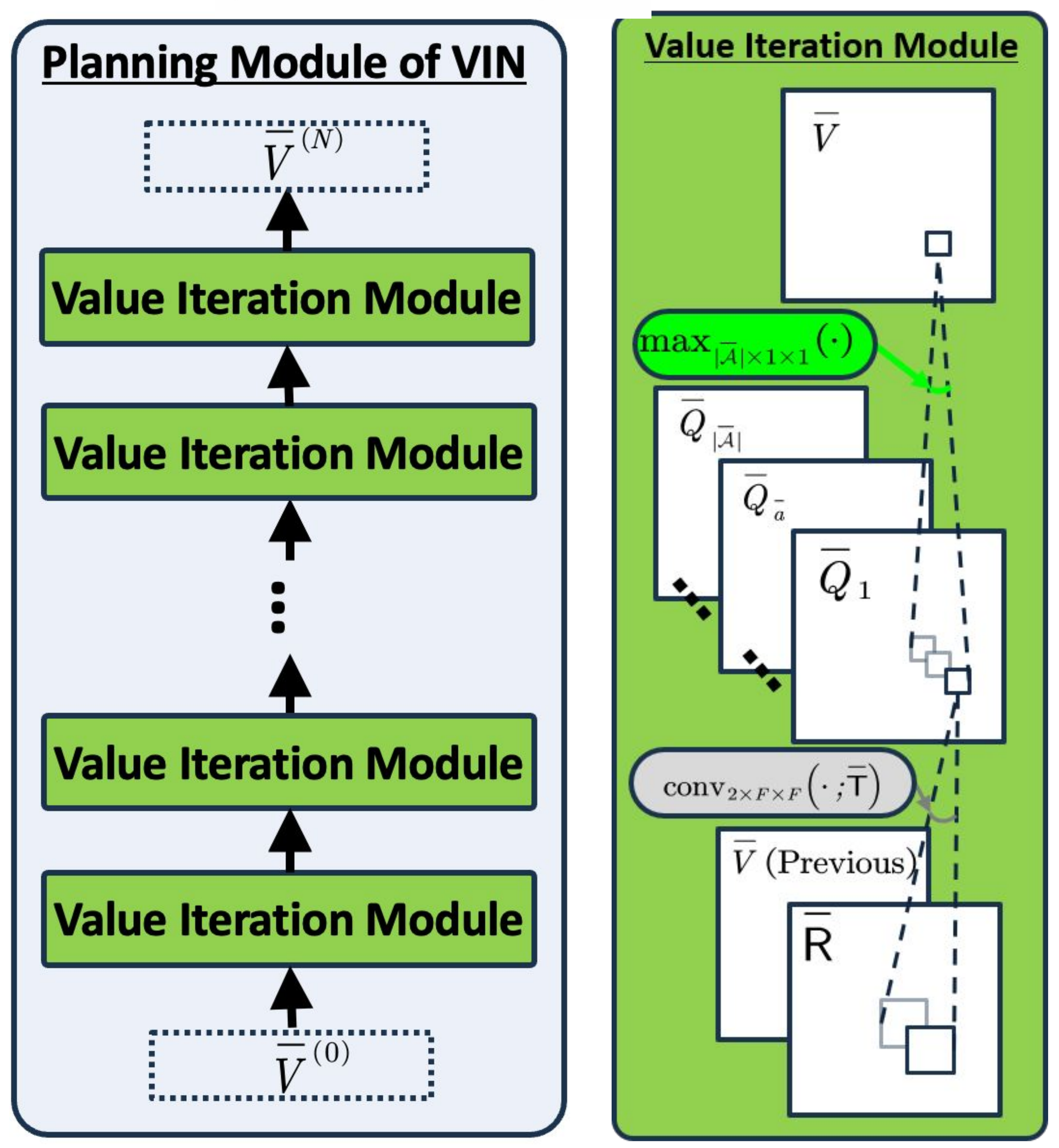
Method

Bellman Optimality/Expectation Operator

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

Planning Module of VIN

$$V^{(n+1)} = \mathcal{B}V^{(n)}$$



Value Iteration Module:

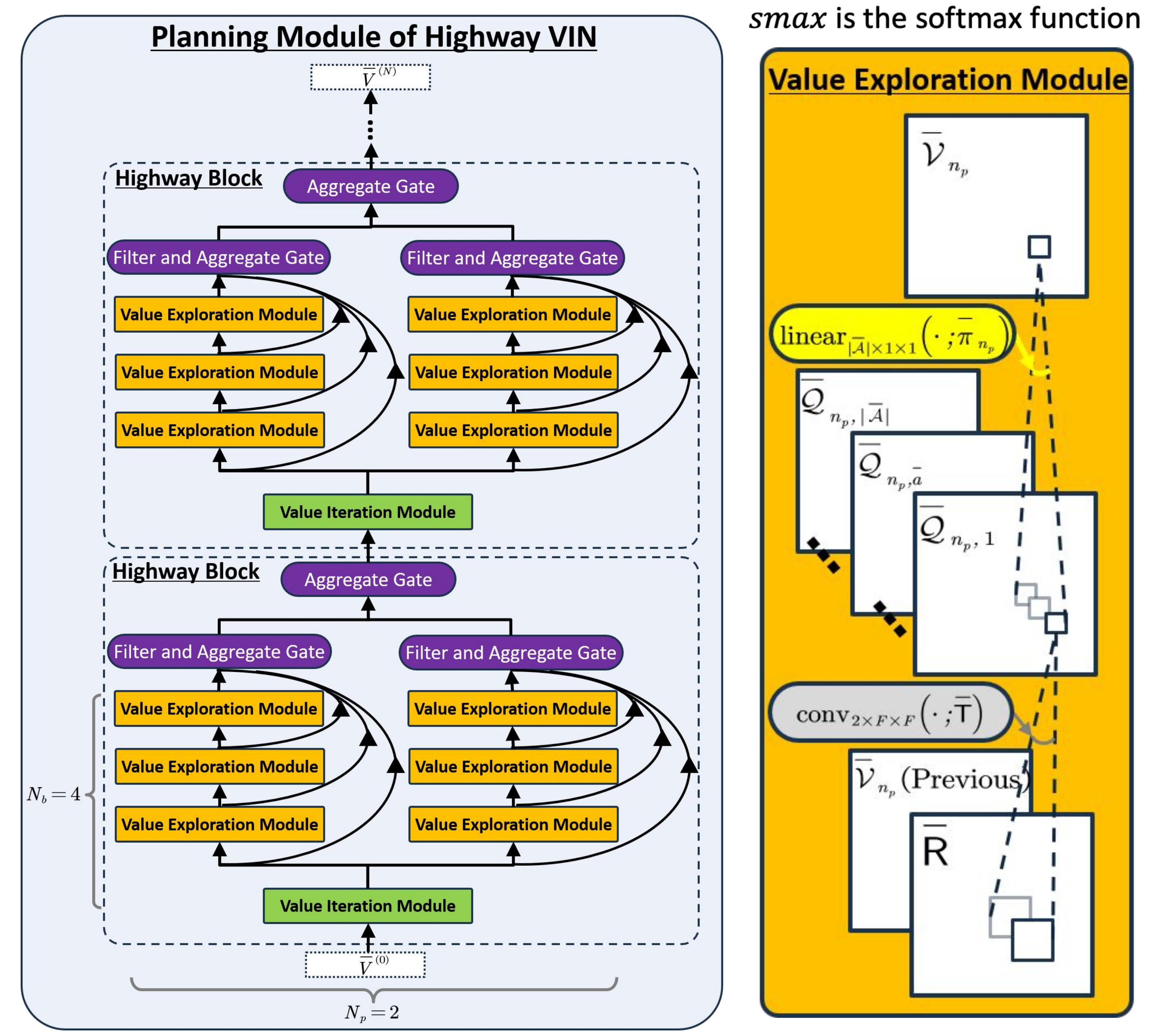
(1) $$\overline{Q}^{(n)}_{\overline{a},i,j} = \sum_{i',j'} \left(\overline{\mathsf{T}}_{\overline{a},i',j'}\overline{\mathsf{R}}_{i-i',j-j'} + \overline{\mathsf{T}}_{\overline{a},i',j'}\overline{V}^{(n-1)}_{i-i',j-j'}\right)$$

(2) $$\overline{V}^{(n)}_{i,j} = \max_{\overline{a}} \overline{Q}^{(n)}_{\overline{a},i,j}$$

Planning Module of Highway VIN

$$V^{(n+1)} = \underset{\pi\in\Pi}{smax}^{\tilde{\alpha}}\ \underset{n\in\mathcal{N}}{smax}^{\alpha} \max\left\{(\mathcal{B}^\pi)^{\circ(N_b-1)}\mathcal{B}V^{(n)}, \mathcal{B}V^{(n)}\right\}$$

smax is the softmax function



Value Exploration Module:

(1) $$\overline{Q}^{(n+n_b)}_{\overline{\pi},\overline{a},i,j} = \sum_{i',j'} \left(\overline{\mathsf{T}}_{\overline{a},i',j'}\overline{\mathsf{R}}_{i-i',j-j'} + \overline{\mathsf{T}}_{\overline{a},i',j'}\overline{V}^{(n+n_b-1)}_{\overline{\pi},i-i',j-j'}\right)$$

(2) $$\overline{V}^{(n+n_b)}_{n_p,i,j} = \sum_{\overline{a}} \overline{\pi}^{(n+n_b)}_{n_p,\overline{a},i,j}\overline{Q}^{(n+n_b)}_{n_p,\overline{a},i,j}$$

where

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

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

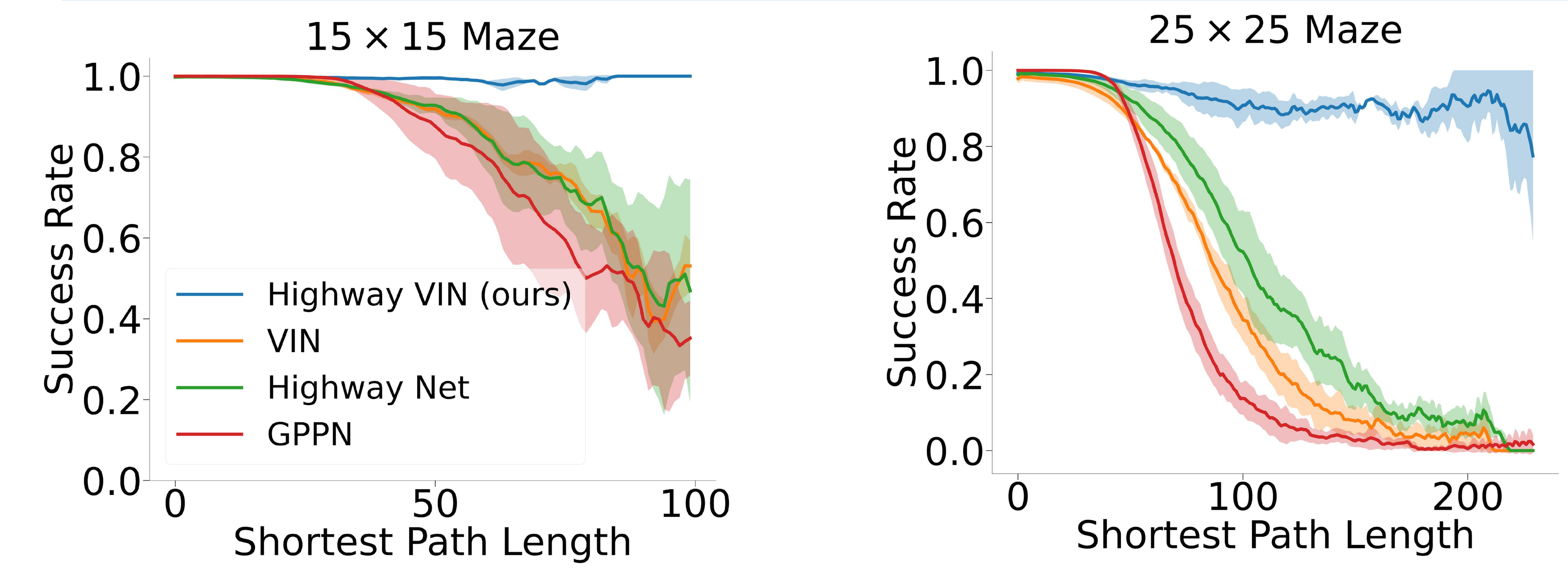
Aggregate Gate: **Filter Gate:**

$$\overline{V}^{(n+N_b)}_{i,j} = \sum_{n_p=1}^{N_p} \widetilde{\mathsf{A}}^{(n+N_b)}_{n_p,i,j} \underbrace{\sum_{n_b=1}^{N_b} \mathsf{A}^{(n+n_b)}_{n_p,i,j} \overbrace{\max\left\{\overline{V}^{(n+n_b)}_{n_p,i,j}, \overline{V}^{(n+1)}_{i,j}\right\}}^{\overline{V}'^{(n+n_b)}_{n_p,i,j}}}_{\overline{V}''^{(n+N_b)}_{n_p,i,j}}$$

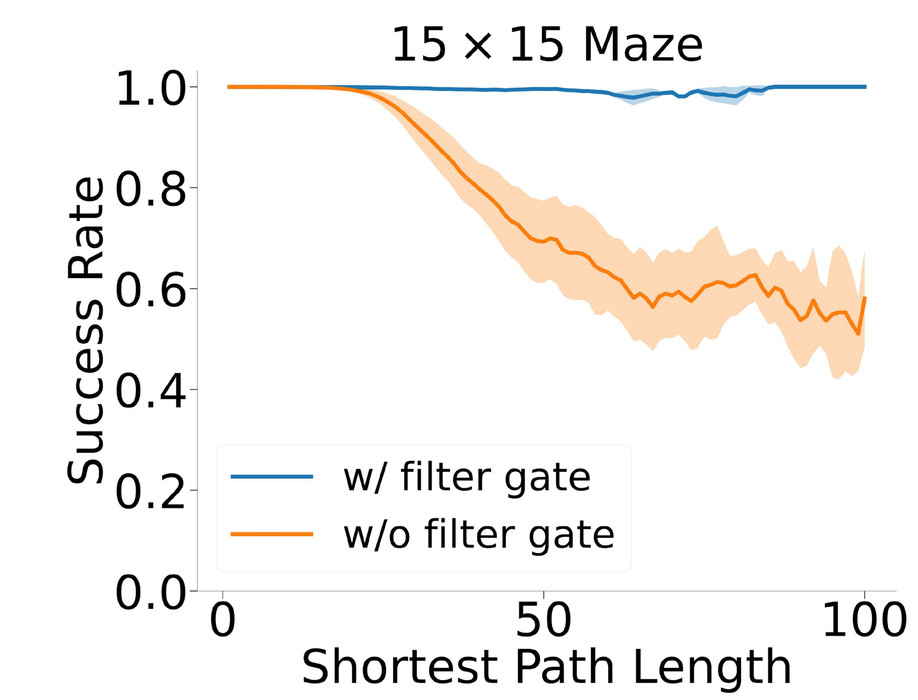
where

$$\widetilde{\mathsf{A}}^{(n+N_b)}_{n_p,i,j} = \frac{\exp\left(\alpha_{\widetilde{\mathsf{A}}}\overline{V}''^{(n+N_b)}_{n_p,i,j}\right)}{\sum_{n_p'}\exp\left(\alpha_{\widetilde{\mathsf{A}}}\overline{V}''^{(n+N_b)}_{n_p',i,j}\right)} \qquad \mathsf{A}^{(n+n_b)}_{n_p,i,j} = \frac{\exp\left(\alpha_{\mathsf{A}}\overline{V}'^{(n+n_b)}_{n_p,i,j}\right)}{\sum_{n_b'}\exp\left(\alpha_{\mathsf{A}}\overline{V}'^{(n+n_b')}_{n_p,i,j}\right)}$$

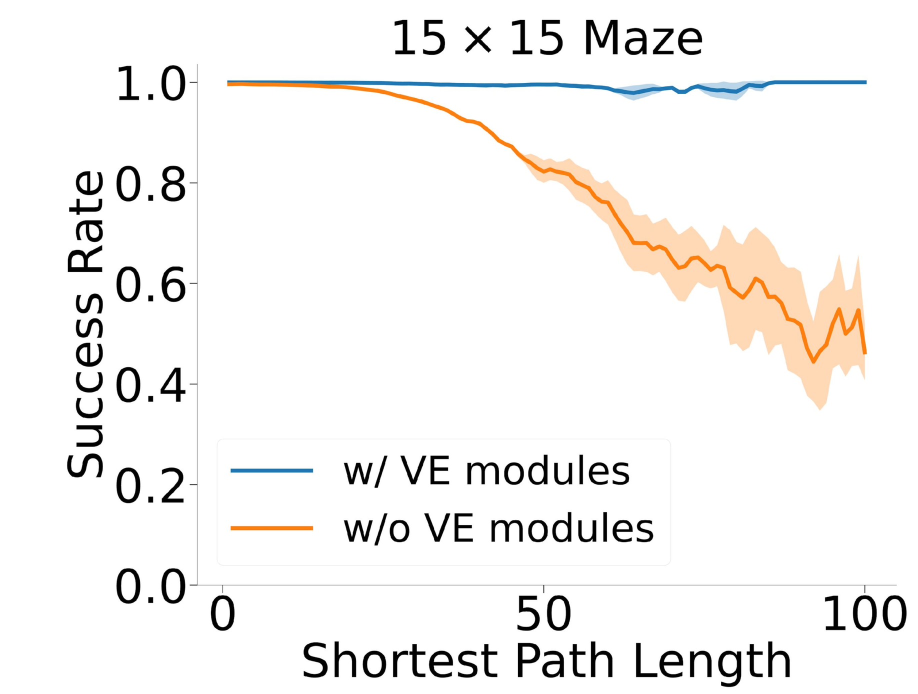
Experiments



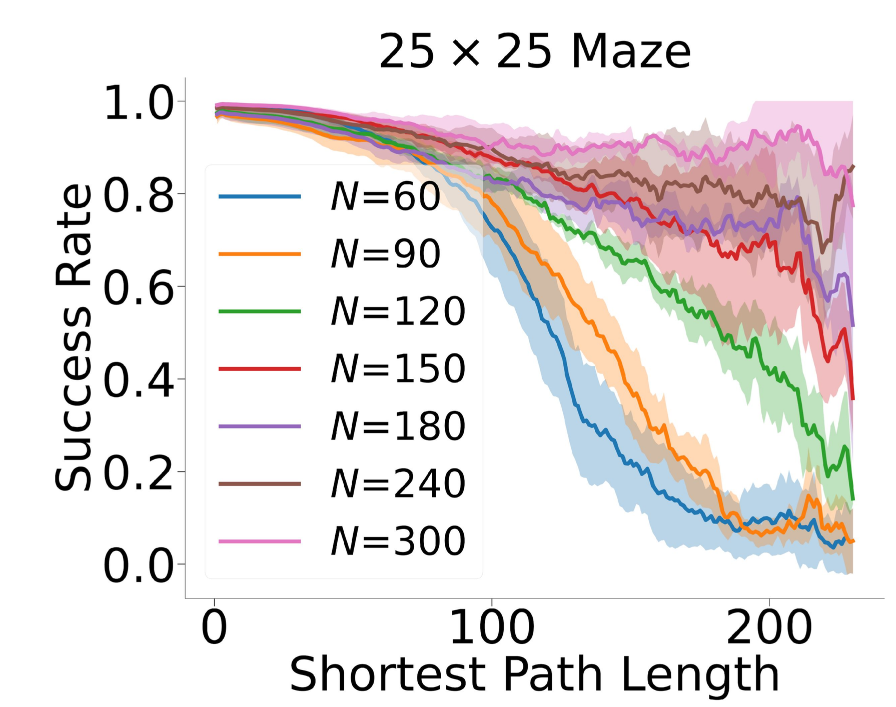
Ablation Studies



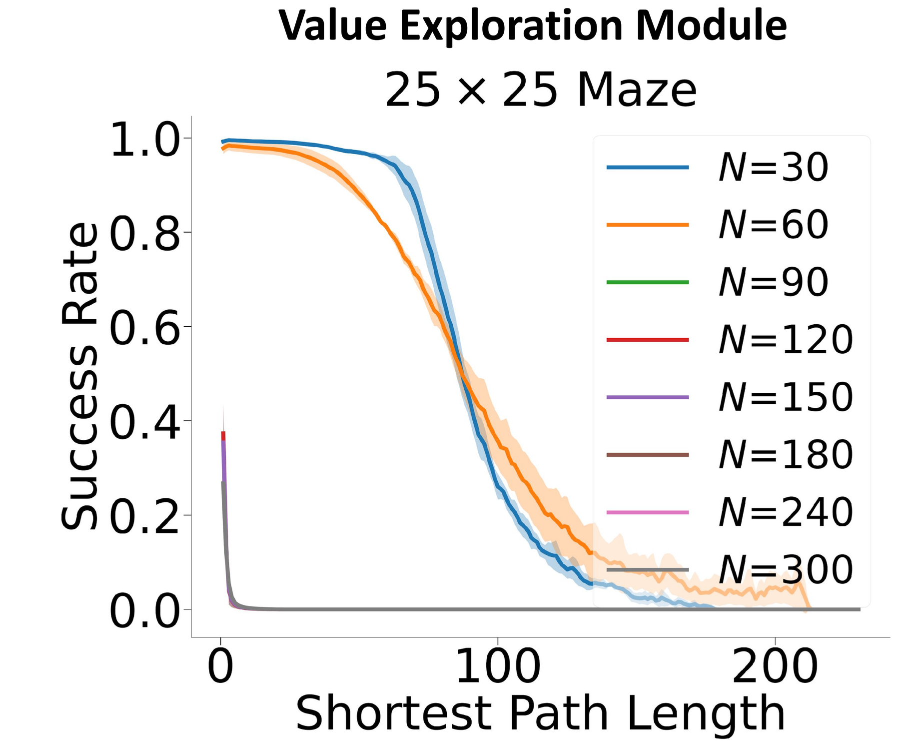
Filter Gate



Value Exploration Module



Depths of Highway VIN (ours)



Depths of VIN

[1] Tamar A, Wu Y, Thomas G, Levine S, & Abbeel P. Value iteration networks[J]. Advances in neural information processing systems, 2016, 29.

[2] Wang Y, Strupl M, Faccio F, Wu Q, Liu H, Grudzien M, Tan X, and Schmidhuber J. Highway reinforcement learning[J]. arXiv preprint arXiv:2405.18289, 2024.