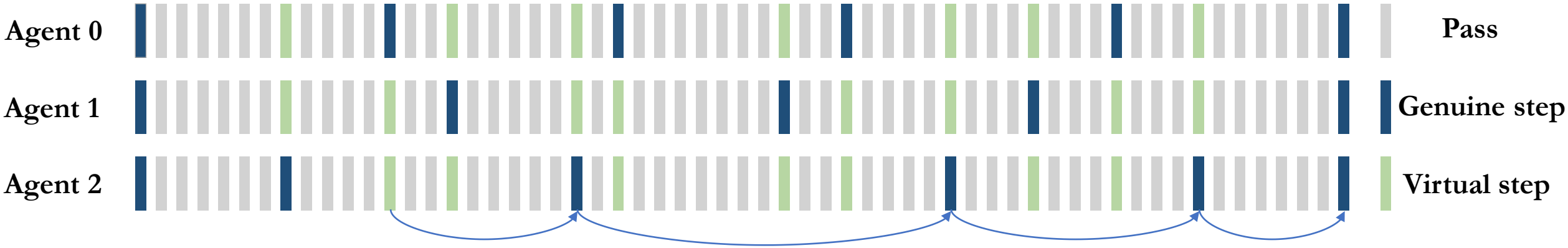

Update **state value function** of each agent, according to

$$L^{m}_{Encoder}(\varphi) = \frac{1}{T_m}\sum_{t=0}^{T_m-1}\left[R(o_t, a_t) + \gamma V_{\bar{\varphi}}\left(o^{im}_{t+1}\right) - V_{\varphi}\left(o^{im}_{t}\right)\right]^2$$ on **genuine step trajectories**.

Update **policy function**, get the recomputed advantage using GAE trick on the trajectories of blue arrow.