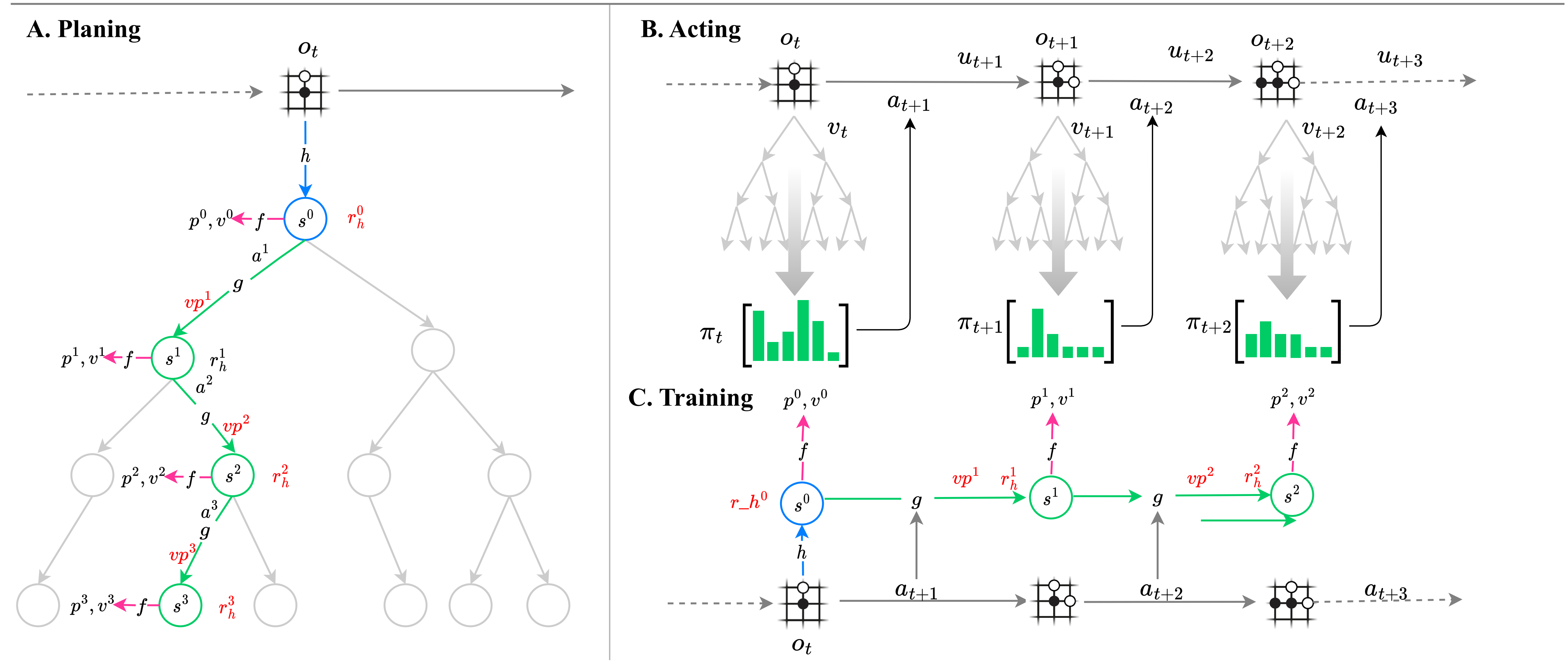
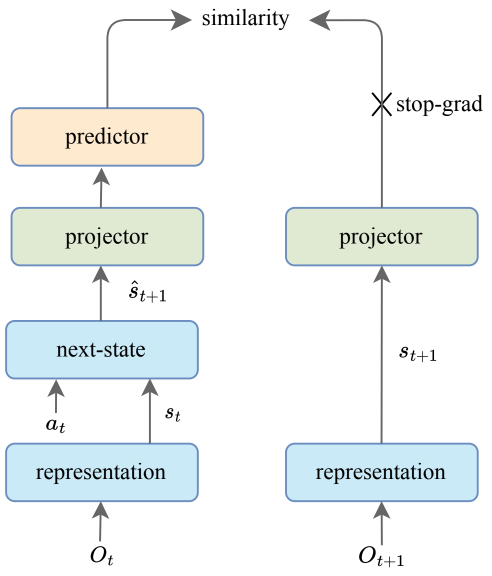
EfficientZero: Self-supervised coonsistency loss, Value prefix, Off-policy correction



D. Network and Loss



The self-supervised consistency loss.

$$l_t(\theta) = \sum_{k=0}^{K} l^r\left(vp_{t+k}, vp_t^k\right) + l^v\left(z_{t+k}, v_t^k\right) + l^p\left(\pi_{t+k}, \mathbf{p}_t^k\right) + l^{similarity}(s_{t+1}, \hat{s}_{t+1}) + c\|\theta\|^2$$

where, $z_t = u_t + \gamma u_{t+1} + \ldots + \gamma^{l-1} u_{t+l-1} + \gamma^l v_{t+l}^{MCTS}$, $l <= k$, v_{t+l}^{MCTS} is reanalyzed MCTS root value.

Reference: MuZero target is $z_t = u_t + \gamma u_{t+1} + \ldots + \gamma^{n-1} u_{t+k-1} + \gamma^k v_{t+k}$.

NOTE：g^{rnn} (abbreviated as g in the figure) is the *recurrent* dynamics network,
vp is value prefix, r_h is reward hidden state.