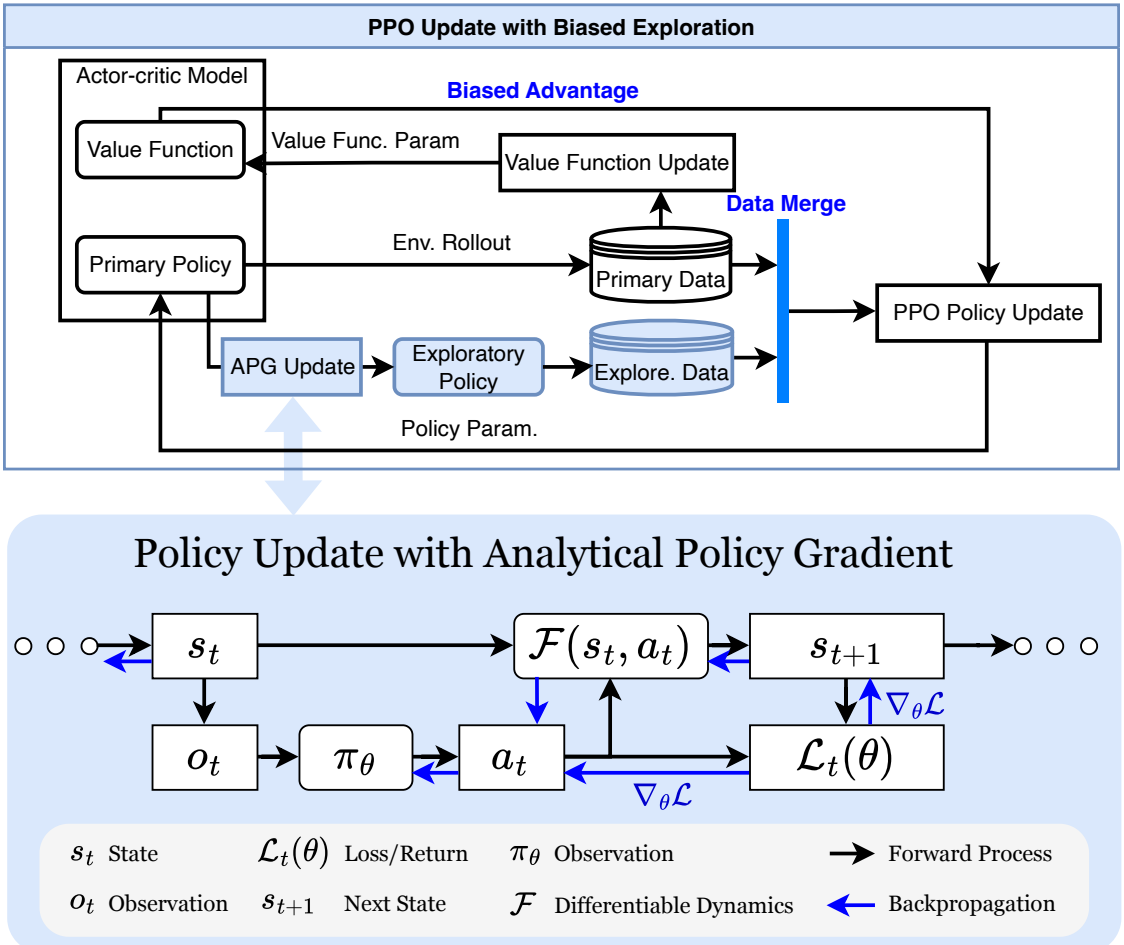
PPO Update with Biased Exploration
Actor-critic Model
Biased Advantage
Value Function
Value Func. Param
Value Function Update
Data Merge
Primary Policy
Env. Rollout
Primary Data
PPO Policy Update
APG Update
Exploratory Policy
Explore. Data
Policy Param.
Policy Update with Analytical Policy Gradient
s_t
$\mathcal{F}(s_t, a_t)$
s_{t+1}
$\nabla_\theta \mathcal{L}$
o_t
π_θ
a_t
$\mathcal{L}_t(\theta)$
$\nabla_\theta \mathcal{L}$
s_t State
$\mathcal{L}_t(\theta)$ Loss/Return
π_θ Observation
Forward Process
o_t Observation
s_{t+1} Next State
$\mathcal{F}$ Differentiable Dynamics
Backpropagation