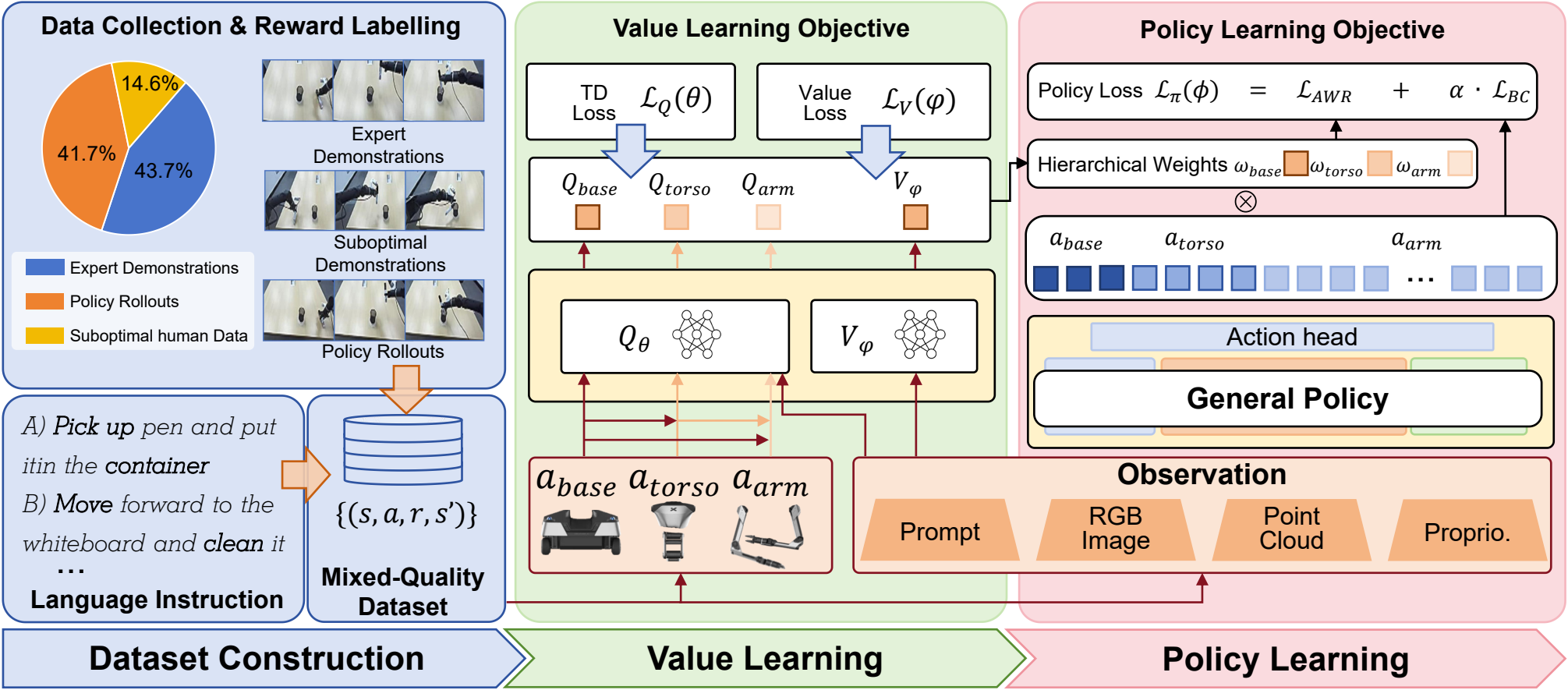
Data Collection & Reward Labelling
14.6%
41.7%
43.7%
Expert Demonstrations
Suboptimal Demonstrations
Policy Rollouts
Expert Demonstrations
Policy Rollouts
Suboptimal human Data
Expert Demonstrations
Suboptimal Demonstrations
Policy Rollouts
A) Pick up pen and put itin the container
B) Move forward to the whiteboard and clean it
...
Language Instruction
$\{(s, a, r, s')\}$
Mixed-Quality Dataset
Value Learning Objective
TD Loss $\mathcal{L}_Q(\theta)$
Value Loss $\mathcal{L}_V(\varphi)$
Q_{base} Q_{torso} Q_{arm} V_φ
Q_θ
V_φ
a_{base} a_{torso} a_{arm}
Policy Learning Objective
Policy Loss $\mathcal{L}_\pi(\phi) = \mathcal{L}_{AWR} + \alpha \cdot \mathcal{L}_{BC}$
Hierarchical Weights ω_{base} ω_{torso} ω_{arm}
a_{base} a_{torso} a_{arm} ...
Action head
General Policy
Observation
Prompt
RGB Image
Point Cloud
Proprio.
Dataset Construction
Value Learning
Policy Learning