Policy Analysis

# 1 · Core hyper-parameters to sweep – what to do

|  |  |  |  |
| --- | --- | --- | --- |
| Hyper-param | How to explore | What to log/observe | Typical signal of “good” |
| Learning rate | Log-scale grid (e.g. 1e-5, 3e-5, 1e-4, 3e-4). Keep other params fixed. | Final success %, time-to-80 % success, KL divergence stability. | Steady increase in return with low KL spikes. |
| Clip ε | Couple each LR with ε ∈ {0.05, 0.1, 0.2, 0.3}. | Policy-value loss ratio, KL, success %. | Mid-range ε often gives fastest but stable learning. |
| GAE λ | 0.90–0.99 in steps of 0.03. | Variance of advantage estimates, learning-curve smoothness. | Higher λ reduces variance but slows response to new rewards. |
| Discount γ | 0.95, 0.98, 0.99. | Long-horizon reward propagation (look at early-episode returns). | Higher γ if task requires long sequences; too high can destabilise. |
| Entropy coef | 0, 1e-3, 1e-2. | Entropy vs success %. | Small coef fights premature convergence. |
| Value-loss coef | 0.5 vs 1.0. | Value-loss magnitude, return. | Match ratio so value and policy losses are similar scale. |
| Batch/mini-batch | Fix #steps → vary mini-batch counts (4, 8, 16). | Sample utilisation, wall-clock time. | Too small → noisy, too large → slow updates. |
| Target KL | 0.01, 0.02, 0.05; early-stop when exceeded. | Number of epochs per update, stability. | Helps avoid catastrophic policy jumps. |

# 2 · Environment & control knobs

|  |  |  |
| --- | --- | --- |
| Knob | Action | Observation focus |
| Action-space scaling | Try torque gain multipliers 0.5×, 1×, 2×. | Avg torque, overshoot, success %. |
| Control frequency | Run env at 20, 50, 100 Hz. | Sample efficiency vs real-time cost. |
| Observation noise / domain randomisation | Add Gaussian noise σ ∈ {0, 0.01, 0.05}. | Robustness curve (success % vs σ). |
| Reward weight tuning | Sweep weights on energy & joint-limit penalties. | Trade-off curves: success vs energy, success vs collisions. |
| Reset-state diversity | Randomise starting joint angles & object pose. | Generalisation test success on unseen initial states. |

# 3 · Metrics to track/plot

Each is a per-run time-series + final mean ± CI

* Episodic return – primary learning curve.
* Success rate – binary completed-task metric.
* Median steps-to-success – efficiency.
* Mean joint torque & power – mechanical stress.
* Policy entropy – exploration vs convergence.
* Target KL per epoch – stability diagnostic.
* Safety events (collisions, joint-limit hits) – practical deployability.

# 4 · Analyses to include

|  |  |  |
| --- | --- | --- |
| Analysis | How to create | Insight delivered |
| Learning curves | Overlay top & worst runs on one plot. | Shows variability & best-run speed. |
| Sensitivity heat-map | 2-D grid (LR × Clip) coloured by success %. | Identifies sweet-spot region. |
| Ablation study | Retrain with entropy=0, no energy penalty, etc. | Quantifies contribution of each term. |
| Generalisation test | Evaluate best policy on unseen payload masses/start poses. | Demonstrates robustness. |
| Sample-efficiency vs baseline | Compare PPO to SAC, DDPG on return vs steps. | Justifies algorithm choice. |
| Noise robustness curve | Success % vs injected obs-noise σ. | Safety under sensing errors. |

# 5 · Tables to include

|  |  |
| --- | --- |
| Table ID | Build it like this |
| T1 Hyper-param grid | One row per run, columns = HP + final success % + best-avg return. Sort by success %. |
| T2 Best-of-N comparison | Rows = algorithm variants; columns = Success %, Steps-to-solve, Avg torque, Collision rate. |
| T3 Ablation results | Rows = variant (e.g. “no entropy”); columns = what’s removed, success %, return, median steps. |

# 6 · One-page figure set

Generate once the sweeps finish; embed in report with short captions.

* Learning-curve overlay – x-axis env-steps, y-axis return.
* LR × Clip heat-map – use seaborn/plt imshow, annotate cells.
* Bar chart of generalisation – x-axis payload categories, y-axis success %.