Skepthical review: *Accelerating Critic Learning via Lyapunov-Structured Value Functions for Reinforcement Learning*

Summary

This paper tests whether adding control-theoretic structure to the critic can accelerate value-function learning in deep RL for continuous control. The critic is parameterized as a residual around an analytic energy/Lyapunov candidate, $V(s) = \Phi(s) + f_\theta(s)$, where Φ is fixed (pendulum mechanical energy) and f_θ is learned (Sec. 2.1, Sec. 2.3.2). The environment reward is replaced with the one-step decrease in Φ: $R_t = \Phi(s_t) - \Phi(s_{t+1})$ (Sec. 2.1). Using PPO on Gymnasium Pendulum-v1, the paper compares a standard critic (Condition A) vs. the Lyapunov-structured critic (Condition B) over 100,000 steps and 5 seeds (Sec. 2.2–2.4). Empirically, the structured critic yields much lower critic loss to GAE targets and value estimates closer to Φ (Sec. 3.2–3.3), but these critic improvements do not translate into better policy performance: both conditions achieve similar (poor) returns and near-zero upright stability (Sec. 3.1, Sec. 3.4). Overall, the paper is clearly written and the negative result is valuable, but the current setup strongly couples the method to a potentially misaligned reward and narrow experimentation, limiting the conclusions about when structured critics help end-to-end RL.

Strengths

- Simple, implementable critic prior: the decomposition $V(s) = \Phi(s) + f_\theta(s)$ is clear and easy to reproduce/extend (Introduction, Sec. 2.3.2).
- Transparent empirical comparison (A vs. B) with multiple seeds and several critic-centric metrics (Sec. 2.2–2.4, Sec. 3.2–3.3).
- The heatmap-style value-function visualizations (Fig. 2) are a useful sanity check that the structured critic meaningfully changes what is learned.
- Honest reporting of the main finding: faster critic convergence does not yield better control under the tested setup (Sec. 3.4, Sec. 4).
- Residual remains non-trivial (not merely copying Φ), supporting the claim that the method refines rather than hard-codes the prior (Sec. 3.3).

Major issues

1. **Reward design likely misaligns with swing-up and is not theoretically consistent with discounting; this undermines interpretation of the negative policy result (Sec. 2.1, Eq. (2); Sec. 2.2; Sec. 3.1; Sec. 3.4).** With $R_t = \Phi(s_t) - \Phi(s_{t+1})$, the (undiscounted) return largely telescopes to a terminal energy difference, and with $\gamma = 0.99$ it does not telescope at all, so Φ is not clearly an approxi-

mation to the discounted value. Moreover, swing-up typically requires temporarily increasing energy; rewarding instantaneous energy decrease can directly discourage that behavior, which is consistent with the near-zero upright stability reported.

Recommendation: Strengthen Sec. 2.1 and Sec. 4 with (i) an explicit derivation/discussion of the relationship between Eq. (2), discounting, and potential-based shaping (contrast with the standard shaping term $F(s, a, s') = \gamma\Phi(s') - \Phi(s)$), and (ii) a behavioral analysis of what Eq. (2) incentivizes from the downward state vs. near-upright. To disentangle reward from critic structure, add a key ablation: keep the environment's default Pendulum-v1 reward (or a known swing-up shaping reward), and change only the critic parameterization $V = \Phi + f_\theta$. If resources are limited, at minimum reframe conclusions as specific to the Lyapunov-decrease reward rather than to structured critics in general.

2. **The core empirical claim ("accelerating critic learning") is evaluated mainly via MSE to GAE targets and proximity to Φ, both of which can be confounded by parameterization/scale and do not directly measure value accuracy for the control problem (Sec. 2.2, Sec. 2.4, Sec. 3.2–3.3).** In Condition B the network learns only a residual, which may have smaller magnitude/simpler structure, making early MSE reductions partly an optimization-conditioning effect rather than evidence of more accurate returns. Also, $\mathbf{MSE}(V, \Phi)$ is expected to improve when Φ is explicitly added to the function class.

   Recommendation: Augment Sec. 3.2–3.3 with value-quality measures that are less self-referential: (1) report MSE between $V(s)$ and Monte Carlo returns estimated from fixed policies (e.g., random policy and snapshots of the learned policy) on held-out rollouts; (2) report explained variance of the value function and/or calibration plots of predicted value vs. empirical returns; (3) for Condition B, also report the residual regression quality, e.g., MSE of $f_\theta(s)$ to the residual target $G_t - \Phi(s_t)$, and provide a scale comparison of targets (variance of G_t vs. $G_t - \Phi$). Clearly qualify in Sec. 3.2/4 that current metrics primarily show easier fitting under the chosen reward/prior, not necessarily better task-relevant value estimation.

3. **The paper's central "critic improves but policy doesn't" conclusion is plausible but currently under-diagnosed; actor–critic interaction and PPO constraints are not analyzed (Sec. 3.1, Sec. 3.4, Sec. 4).** Without actor diagnostics, it is unclear whether the actor is bottlenecked by exploration, advantage quality/scale, PPO clipping/KL constraints, or by the reward itself.

   Recommendation: Add PPO/actor diagnostics in Sec. 3.1 and Sec. 3.4: track advantage distribution statistics (mean/variance, fraction positive), policy gradient norm, entropy/action-std evolution, clip fraction, and KL divergence per update for Conditions A/B. Include a small set of representative evaluation trajectories (angle/velocity/energy vs. time) to show what behaviors are being learned under the Lyapunov re-

ward. Use these to sharpen the discussion in Sec. 4 (e.g., whether the critic signal is improved but the effective policy update is unchanged due to clipping or low exploration).

4. **Experimental scope is too narrow to support broad claims about on-policy RL or structured critics (Sec. 2–4).** Only Pendulum-v1, PPO, one reward design, one training budget (100,000 steps), and 5 seeds are used; the method might matter more in settings where critic quality more directly impacts learning (e.g., off-policy) or with longer horizons.

   Recommendation: Broaden evaluation where feasible: (i) extend training budgets (e.g., 300,000–1,000,000+) to test whether the critic advantage eventually yields policy improvements; (ii) add at least one additional environment with an interpretable Φ (e.g., cart-pole variants, simple LQR, double integrator) and/or one off-policy baseline (SAC/TD3) where improved critic bootstrapping/sample reuse could plausibly translate to policy gains. If expansion is not possible, tighten Sec. 4 to explicitly limit conclusions to Pendulum-v1 + PPO + the specific Lyapunov-decrease reward + 100,000-step regime.

5. **Key implementation details affecting the critic targets and reported loss reductions are under-specified, reducing reproducibility and making it hard to interpret the magnitude of improvements (Sec. 2.2–2.4).** Ambiguities include: how GAE/TD targets are computed with the modified reward, how bootstrapping is handled at truncation vs. termination (Pendulum episodes are short), whether any reward/value normalization is used, and the exact optimizer/value-loss/entropy coefficients and clipping/gradient settings.

   Recommendation: In Sec. 2.2–2.3.2, explicitly write the target computation used (TD residuals, GAE equations), stating whether bootstrapping uses the full $V(s) = \Phi(s) + f_\theta(s)$ and how terminal vs. truncated transitions are treated. In Sec. 2.2 and Sec. 2.4, list PPO hyperparameters comprehensively (optimizer + betas, lr schedule, value/entropy coefficients, grad clipping, observation/reward normalization, network architecture and whether actor/critic share layers). For the 100×100 grid evaluation (Sec. 2.4, Sec. 3.3), specify exact θ and $\dot{\theta}$ ranges, mapping to $[\cos\theta, \sin\theta, \dot{\theta}]$, and averaging procedure.

Minor issues

1. Policy-performance evaluation is too thin to support the "no stable policy" characterization, and episodic return under the shaped reward is not interpreted behaviorally (Sec. 3.1, Sec. 3.4). Upright stability is reported only as an aggregate fraction with a single threshold $|\theta| < 0.1$, with limited context for what returns like -0.78 mean under Eq. (2).

Recommendation: In Sec. 3.1 and Sec. 3.4, add complementary metrics: average $|\theta|$, average $|\dot{\theta}|$, average $\Phi(s)$, time-to-upright (first hitting time to $|\theta| < \epsilon$ and $|\dot{\theta}| < \epsilon$), and/or a success rate over episodes. Plot upright stability over training (or per-seed ranges) to show whether any runs briefly succeed. Add a short interpretation mapping return magnitudes under Eq. (2) to typical behaviors (e.g., energy dissipation without swing-up vs. near-upright stabilization).

2. Terminology and conceptual framing around "Lyapunov function" and the Lyapunov–value-function relationship is overstated given what is proven/used (Sec. 1, Sec. 2.1, Sec. 4). Eq. (1) is energy-like and positive definite around the equilibrium, but the paper does not establish decrease along the learned closed-loop dynamics; also Φ is not generally equal to the discounted value under Eq. (2).

   Recommendation: Adjust wording to "energy-like candidate Lyapunov function" or "control-theoretic prior" unless a decrease condition is established. Add a brief clarification in Sec. 1/Sec. 2.1 about when value functions resemble Lyapunov functions (specific costs/dynamics/undiscounted settings) and cite relevant control/RL references.

3. Related work is currently too limited, making novelty and positioning unclear (Sec. 1–2). The paper touches Lyapunov ideas, residualization, analytic priors, and safe/stable RL, but does not connect to established lines of work.

   Recommendation: Add a related-work subsection (end of Sec. 1 or new Sec. 2.5) covering: Lyapunov-based RL/safe RL, potential-based reward shaping, residual/value decomposition approaches, and methods incorporating analytic models/priors into critics. Explicitly state what is new here (e.g., isolating critic optimization effects and reporting the critic–policy decoupling under a simple prior).

4. Figures need improvements for interpretability and to support claims (Fig. 1–2; Sec. 3.1–3.3). Fig. 1 has naming inconsistencies (Direct/Structured vs. Condition A/B) and limited quantitative summary; Fig. 2 lacks shared color scale across comparable panels and has legibility issues (axes/ranges/panel labels).

   Recommendation: Standardize nomenclature across text/legend/captions (explicitly map A/B everywhere). For Fig. 1, include end-of-training summary (mean ± CI over last N steps) in caption and clarify shading semantics/smoothing. For Fig. 2, use identical color limits for Φ, V_A, V_B with a shared colorbar (report min/max), add clear axis labels/ranges and panel labels (a–d), and consider adding an error map (e.g., $V - \Phi$) to highlight where the residual matters.

5. Aggregation/variability reporting is incomplete for some scalar summaries (Sec. 3.2–3.3, Table 1). Some metrics appear as means without clear averaging windows and without uncertainty.

Recommendation: For Table 1 and any scalar metrics (early/overall loss, $\mathbf{MSE}(V, \Phi)$), report mean ± std (or CI) across seeds and define precisely the time window/aggregation (e.g., first $\mathbf{10{,}000}$ steps; average over all updates; last $\mathbf{10{,}000}$ steps).

6. State/angle handling is slightly ambiguous: Φ is written as a function of $\theta, \dot{\theta}$ while the environment observation is $[\cos\theta, \sin\theta, \dot{\theta}]$; it is not fully explicit that the same $\theta = \mathbf{atan2}(\sin\theta, \cos\theta)$ mapping is used in reward computation vs. evaluation (Sec. 2.1, Sec. 2.4).

   Recommendation: State explicitly in Sec. 2.1 (reward computation) how θ is obtained from the observation and confirm consistency between training-time reward computation and evaluation-time plotting/grid evaluation.

Very minor issues

1. Formatting/cross-reference issues reduce readability: unresolved "Figure ??"/"Table ??" references, stray heading markers (e.g., leading "#"), HTML escapes (e.g., "<"), and inconsistent notation for $\mathbf{100 \times 100}$ (Sec. 2.4, Sec. 3.2–3.4).

   Recommendation: Fix LaTeX/build issues: resolve all cross-references, standardize section heading formatting, remove HTML escapes, and unify notation (e.g., consistently use $\mathbf{100 \times 100}$).

2. Minor style/presentation polish: inconsistent spacing around percentages and inline math, and occasional informal phrasing (Sec. 3–4).

   Recommendation: Proofread for consistent typography (e.g., "86\%", "mean~±~standard deviation") and slightly more formal phrasing where needed.

3. Nonstandard/placeholder author affiliation line ("Anthropic, Gemini~\&~OpenAI servers. Planet Earth.") is inappropriate for an academic submission.

   Recommendation: Replace with standard affiliations (or anonymize properly if double-blind).

Key statements and references

- • **Lyapunov stability theory provides a rigorous method for certifying system stability around an equilibrium point by constructing a scalar Lyapunov function that is positive definite and decreases along all system trajectories, whose existence guarantees convergence to the equilibrium for many physical systems where such analytical Lyapunov functions can be derived from first principles.**
- *Reference(s):* (none)

- • **The Proximal Policy Optimization (PPO) algorithm is an on-policy actor–critic method that uses a clipped surrogate objective with a clipping ratio $\epsilon = 0.2$, Generalized Advantage Estimation with parameter $\lambda = 0.95$, and typically employs separate neural networks for the actor and critic updated from rollouts of collected on-policy data.**
- *Reference(s):* (none)
- • **The Gymnasium Pendulum-v1 environment is a standard continuous-control benchmark in which the task is to swing up an underactuated pendulum and stabilize it in the upright position, with the state commonly represented as $s = [\cos\theta, \sin\theta, \dot{\theta}]$ and the action as a continuous torque applied at the joint.**
- *Reference(s):* (none)

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains a small set of central analytic definitions (a candidate Lyapunov/energy function Φ, a shaped reward defined as its one-step decrease, and a critic value-function decomposition $V = \Phi + \text{residual}$). There are no detailed derivations of PPO/GAE mathematics in the manuscript text, so the audit focuses on internal consistency of these definitions, their relationship to discounted returns, and symbol/definition consistency with the stated state representation.

Checked items

1. ✔ **Candidate Lyapunov/energy function definition** (Eq. (1), Sec. 2.1, p.3)
    - **Claim:** Defines $\Phi(s) = (1 - \cos\theta) + 0.5 \cdot \dot{\theta}^2$ as a Lyapunov/energy-like function relative to upright equilibrium.
    - **Checks:** symbol/definition consistency, sanity/limiting case
    - **Verdict:** PASS; confidence: high; impact: moderate
    - **Assumptions/inputs:** θ is the pendulum angle with $\theta = 0$ at upright., $\dot{\theta}$ is angular velocity., s contains sufficient information to obtain θ and $\dot{\theta}$.
    - **Notes:** Analytically, $\Phi(0,0) = 0$ and near $\theta = 0$, $1 - \cos\theta \approx \theta^2/2$, so Φ is locally positive definite in $(\theta, \dot{\theta})$. No coefficients/units are specified; as written it is dimensionless but internally consistent.
2. ✔ **Reward as one-step decrease in Φ** (Eq. (2), Sec. 2.1, p.3)
    - **Claim:** Defines $R_t = \Phi(s_t) - \Phi(s_{t+1})$, incentivizing energy decrease.
    - **Checks:** algebraic sanity, simple limiting case

- **Verdict:** PASS; confidence: high; impact: moderate
- **Assumptions/inputs:** $\Phi(s_t)$ and $\Phi(s_{t+1})$ are computed consistently from the encoded state.
- **Notes:** If Φ decreases over a transition, $R_t > 0$; if Φ increases, $R_t < 0$. For an undiscounted finite-horizon sum, $\sum_t R_t$ telescopes to $\Phi(s_0) - \Phi(s_T)$, which is mathematically consistent with the definition.

3. ⚠ **Discounted return induced by Eq. (2) vs. Φ baseline** (Sec. 2.2 ($\gamma = 0.99$) + Eq. (2), pp.3–4; mention of "discounted returns" in Sec. 3.2, p.5)

   - **Claim:** Motivates Φ as a strong initialization/prior for predicting "actual discounted returns" used as critic targets.
   - **Checks:** derivation completeness, consistency with discounting
   - **Verdict:** UNCERTAIN; confidence: medium; impact: critical
   - **Assumptions/inputs:** PPO/GAE uses discounted returns with $\gamma = 0.99$ as stated., Critic is trained against GAE-computed return/targets G_t.
   - **Notes:** Given $R_t = \Phi(s_t) - \Phi(s_{t+1})$, the discounted sum $\sum_k \gamma^k R_{t+k}$ generally does not reduce to a simple function $\Phi(s_t)$ (telescoping fails when $\gamma \neq 1$). The paper does not provide the algebra showing under what assumptions Φ approximates the discounted value. This missing analytic step is central to the stated rationale for the structural prior.

4. ✔ **Structured critic decomposition** (Eq. (3), Sec. 2.3.2, p.4)

   - **Claim:** Defines the critic as $V(s) = \Phi(s) + f_\theta(s)$, training only the residual network f_θ.
   - **Checks:** algebraic equivalence, definition consistency
   - **Verdict:** PASS; confidence: high; impact: critical
   - **Assumptions/inputs:** Φ is fixed and non-trainable., Training minimizes MSE between $V(s)$ and return targets (as in Sec. 2.3.1 and Sec. 3.2 description).
   - **Notes:** If the loss is $(V(s) - G_t)^2$ and $V(s) = \Phi(s) + f_\theta(s)$, then optimizing θ is equivalent to regressing $f_\theta(s)$ onto $(G_t - \Phi(s))$, which is internally consistent.

5. ⚠ **State representation vs. θ usage in Φ and stability metric** (Sec. 2.1 (state definition), p.2; Eq. (1), p.3; Upright stability definition, Sec. 2.4 and Sec. 3.4, pp.4 and 6)

   - **Claim:** Uses $s = [\cos\theta, \sin\theta, \dot{\theta}]$ while defining Φ in terms of θ, and evaluates upright condition $|\theta| < 0.1$ with θ recovered via atan2.
   - **Checks:** symbol/definition consistency
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** $\theta = \text{atan2}(\sin\theta, \cos\theta)$ is used wherever θ is needed.

- **Notes:** The paper explicitly mentions atan2 recovery for the upright metric, but does not explicitly state the same recovery is used when computing $\Phi(s)$ for rewards/training and for the state-grid evaluation. This is likely intended, but not stated; the missing clarification prevents a fully rigorous internal consistency check.

6. ✔ **Telescoping property (undiscounted) vs. reported episode return definition** (Sec. 2.4 (episode return = sum of Lyapunov-based rewards), p.4; Eq. (2), p.3)
    - **Claim:** Episode return is the sum of R_t over an episode.
    - **Checks:** algebraic sanity
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** Episode return is undiscounted sum as stated (not γ -discounted)., Episodes terminate after 200 steps.
    - **Notes:** If the episode return is the plain sum $\sum_t R_t$, it telescopes to $\Phi(s_0) - \Phi(s_T)$. The text defines episode return as a sum but does not explicitly say discounted vs. undiscounted; however, the phrasing supports the undiscounted interpretation, making this internally consistent.

Limitations

- Only the content present in the provided PDF text/images was used; PPO/GAE mathematics (e.g., explicit definitions of G_t, advantage estimates, and the critic loss formula) are referenced but not written out, limiting derivation-level verification.
- No environment dynamics equations are provided, so no analytic Lyapunov decrease condition ($d\Phi/dt < 0$ or discrete-time $\Phi(s_{t+1}) - \Phi(s_t) < 0$ under a controller) can be checked from the manuscript itself.
- Several references to figures/tables are unresolved ("??"), reducing traceability but not enabling further symbolic checks.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

All candidate numeric statements that were checkable via arithmetic relationships among reported summary values passed. This includes the claimed $\sim 8\times$ early-loss difference (**5.686** vs **0.734**), the claimed **87%** lower overall loss (**1.057** vs **0.136**), and the claimed **86%** reduction in value-function MSE (**71.98** vs **10.07**). Additional internal-consistency checks (episode/step arithmetic, rollout/update counts, minibatch counts, grid-point count) also matched exactly.

Checked items

1. ✔ **C1__early__loss__ratio__8x** (p.5, Table 1 and §3.2 text)
    - **Claim:** "Condition B achieves an 8-fold lower critic loss during early training (**0.734** vs. **5.686**)"

- **Checks:** ratio_check
- **Verdict:** PASS
- **Notes:** Computed ratio $5.686/0.734 = 7.7466$, within 5% relative tolerance of the claimed 8-fold.

2. ✔ **C2_overall_loss_percent_lower_87** (p.5, Table 1 caption; p.1 Abstract; p.7 Conclusions)
   - **Claim:** "Condition B achieves 87% lower overall loss" (overall mean 1.057 vs. 0.136).
   - **Checks:** percent_change_check
   - **Verdict:** PASS
   - **Notes:** Computed $100 \times (1.057 - 0.136)/1.057 = 87.1334\%$, within 1 percentage point of the claimed 87%.

3. ✔ **C3_value_mse_percent_reduction_86** (p.5, §3.3 text; p.6 Figure 2 caption; p.7 Conclusions)
   - **Claim:** "MSE(V_A, Φ) = 71.98 vs. MSE(V_B, Φ) = 10.07 — an 86% reduction / 86% closer."
   - **Checks:** percent_reduction_check
   - **Verdict:** PASS
   - **Notes:** Computed $100 \times (71.98 - 10.07)/71.98 = 86.0100\%$, matching the claimed 86% within tolerance.

4. ✔ **C4_final_loss_comparison_direction** (p.5, Table 1; §3.2 text)
   - **Claim:** Table 1 reports final training losses: Condition A 0.0010 ± 0.0003 vs Condition B 0.0029 ± 0.0051; verify which is lower and basic plausibility of the claim "Both conditions converge to similarly low final losses."
   - **Checks:** inequality_and_scale_check
   - **Verdict:** PASS
   - **Notes:** Direction check satisfied: 0.0010 (A) $<$ 0.0029 (B). Scale ratios (final/overall and final/early) were all $<! < 1$, supporting 'similarly low' in magnitude (though B is larger than A).

5. ✔ **C5_episode_return_similarity_diff** (p.6, Table 2)
   - **Claim:** Mean episode return: Condition A -0.778 ± 0.168 vs Condition B -0.775 ± 0.167; verify the absolute difference is small compared with reported std.
   - **Checks:** difference_vs_uncertainty_check
   - **Verdict:** PASS
   - **Notes:** Absolute mean difference is 0.003, which is $\sim 0.018\times$ the reported std for each condition (0.168, 0.167), consistent with 'similar returns'.

6. ✔ **C6_upright_fraction_ratio** (p.6, §3.4 and Table 2)
    - **Claim:** Upright stability near zero: 0.0010 ± 0.0018 (A) vs 0.0041 ± 0.0038 (B); verify magnitudes and relative increase.
    - **Checks:** ratio_and_bounds_check
    - **Verdict:** PASS
    - **Notes:** Reported means are within $[0, 1]$, and the computed ratio $0.0041/0.0010 = 4.1$ matches the stated $\sim 4.1\times$ increase.
7. ✔ **C7_steps_per_episode_consistency** (p.3 §2.1; p.4 §2.4)
    - **Claim:** Episodes terminated after 200 timesteps; total training is $100{,}000$ environment steps; verify implied maximum number of episodes (if fully packed) is 500.
    - **Checks:** integer_division_check
    - **Verdict:** PASS
    - **Notes:** $100{,}000/200 = 500$ exactly; remainder 0.
8. ✔ **C8_rollout_steps_vs_total_updates** (p.3 §2.2; p.4 §2.4)
    - **Claim:** Data collected in rollouts of 2048 steps; total training $100{,}000$ steps; verify implied number of full rollouts ≈ 48.8 and thus about 48–49 policy updates.
    - **Checks:** division_check
    - **Verdict:** PASS
    - **Notes:** $100{,}000/2048 = 48.828125$ rollouts; floor/ceil are 48 and 49 with remainder 1696 steps after 48 full rollouts.
9. ✔ **C9_minibatches_per_epoch** (p.3 §2.2)
    - **Claim:** Rollout size 2048 steps, minibatch size 64, 4 epochs; verify minibatches per epoch is $2048/64 = 32$ and total gradient steps per rollout is $32 \times 4 = 128$.
    - **Checks:** divisibility_and_product_check
    - **Verdict:** PASS
    - **Notes:** 2048 is divisible by 64 with remainder 0; minibatches/epoch $= 32$ and total over 4 epochs $= 128$.
10. ✔ **C10_state_grid_size** (p.4 §2.4; p.5 §3.3)
    - **Claim:** Value function evaluated on a 100×100 grid; verify this implies $10{,}000$ grid points.
    - **Checks:** product_check
    - **Verdict:** PASS
    - **Notes:** $100 \times 100 = 10{,}000$ grid points.

Limitations

- Only parsed text from the PDF was available; no underlying experimental logs, model checkpoints, or tabular raw data beyond the reported summary numbers.
- Checks avoid extracting values from plotted figures/heatmaps because pixel-based value extraction is out of scope.
- Several reported metrics (critic losses, MSE on state grid, stability fractions) are outputs of experiments and cannot be recomputed from the PDF alone; only arithmetic relationships among reported numbers can be verified.
- Some statements (residual magnitude as a percent of Φ, learning-curve convergence behavior, policy stability within 100,000 steps, and the detailed computation/definition behind critic-loss MSE vs. GAE returns) remain unverified without underlying data.