Reward Maximization Collapses Generative Diversity: Characterizing and Controlling the Trade-off in Verifiable Procedural Generation

Ali Asaria , Tony Salomone , and Deep Gandhi*

Transformer Lab

Abstract

We study what happens to *generative diversity* when a language model is post-trained with verifiable-reward reinforcement learning (RLVR) to produce solvable artifacts, using procedural Sokoban-level generation as a clean, exactly-verifiable testbed. We find a sharp, reproducible **reward↔diversity phase transition**: as the model learns to maximize a gated, solver-checked reward, it abruptly mode-collapses, generating one or two level templates, with the fraction of distinct valid levels falling from ≈ 1.0 to < 0.05 over a handful of optimization steps. The collapse reproduces across three trust-region objectives (PPO clipping, DAPO `clip-higher`, and the recent DPPO divergence-gating), three seeds, and two model scales. We then show the trade-off is *controllable*: (i) **passively**, by early-stopping at the Pareto knee, and (ii) **actively**, via a diversity-aware reward (an intra-batch novelty bonus) that *induces* additional diversity at high reward: it removes *exact* mode collapse (distinct levels rise $\sim 8\times$) but only modestly restores *structural* variety (pairwise cell-disagreement $0.00 \to 0.03$ vs. 0.39 for an un-collapsed model), suggesting the effect is real but small at our scale and may require larger-scale training to promote. Larger models enjoy a strictly better diversity/reward Pareto frontier. In contrast, *which* trust-region objective is used is **not** a robust lever for diversity on this task: an apparent advantage for DPPO at 1.5B does not survive seeds or scale, a cautionary negative for the assumption that the clipping mechanism governs generative diversity.

1 Introduction

PPO's ratio clipping has had a "second wave" in the LLM era [4, 5], with an active debate about whether its trust-region proxy is mis-specified and how that affects exploration and entropy. DAPO argues the clip range is too tight and loosens the upper bound (`clip-higher`) [8]; sequence-level variants change the ratio granularity [9]; and DPPO argues the per-token ratio is the wrong gating variable and replaces it with a direct policy-divergence mask [3]. A recurring claim is that better trust-region handling preserves *exploration*, which, for a *generator*, should manifest as output *diversity*.

We test this directly in a setting where diversity is the quantity of interest and where reward is *exactly* verifiable: generating solvable Sokoban levels of a target difficulty. This turns the abstract "clipping affects entropy" discussion into a concrete, measurable question about the diversity of generated artifacts, and lets us study it without ever rewarding diversity. Our contributions:

*Corresponding author: `deep@lab.cloud`

1. A characterization of a sharp **reward↔diversity phase transition** in verifiable procedural-generation RL, reproducible across objectives, seeds, and scales.
2. Control of the trade-off: early-stopping at the Pareto knee (passive, recovers genuine diversity) and a diversity-aware reward (active, but a small effect at our scale: it removes exact collapse yet only modestly restores structural variety).
3. A scaling result: larger models have a **dominating** diversity/reward frontier.
4. An honest **negative**: the trust-region objective (PPO/DAPO/DPPO) is not a robust diversity lever here; a 1.5B effect does not replicate at 3B.

2 Related Work

Our objectives are GRPO [5] with three trust-region variants: symmetric PPO clipping [4], DAPO `clip-higher` [8], and DPPO's divergence-gated mask [3]; we also consider token- vs. sequence-level ratios [9]. Procedural content generation via RL has a long line of work [2]; we instead drive an LLM generator with a verifiable reward. Boxoban [1] supplies format-priming levels. Models are Qwen2.5-Instruct [7].

3 Method

Task and verifier. The model emits a 10×10 ASCII Sokoban grid with exactly four boxes/goals and one player. An exact A* solver (BFS player-reachability; admissible min-cost box→goal assignment; static dead-square pruning) decides solvability and returns the optimal push-count, our difficulty proxy. The solver is the fixed evaluator; it solved 120/120 sampled Boxoban levels at $\approx$0.08 s each, cheap enough to use in the RL loop.

Reward (gated). A generation earns reward 1 iff it meets the exact spec *and* is solvable *and* its optimal push-count is in $[12, 30]$; partial credit (0.3 solvable, 0.1 valid-spec) shapes early learning. **Diversity is never part of the gated reward**, so any change in generated diversity is attributable to training, not to the reward. Tightening the reward to the exact 10×10/4-box spec was necessary: a looser solver-only reward was gamed by emitting *smaller* solvable boards.

Objectives. All arms use GRPO. `vanilla` uses symmetric clipping (ϵ=0.2); `clip-higher` decouples ϵ_{low}=0.2, ϵ_{high}=0.28 [8]; `DPPO` replaces the clip with a binary mask gated on the total-variation divergence $D_t = |\mu(a_t) - \pi(a_t)|$ between the rollout policy μ and current policy π, blocking only updates that push the ratio further from 1 when $D_t > \delta$ (δ=0.15) [3].

Diversity metrics. We report, over all spec-valid generations, the distinct rate under a dihedral-(D_4)-canonical form (so rotations/reflections are not counted as new), canonical-form entropy, mean pairwise cell-disagreement, and character n-gram diversity. *Distinct@valid* is the headline metric; distinct-over-in-band saturates at 1.0 and is uninformative.

Diversity-aware reward (active method). We add an intra-batch novelty bonus: a spec-valid generation receives $+\lambda/\,c$ where c is the count of its canonical form in the batch (λ=0.5). Unique levels get the full bonus; repeated templates get a shrinking share, so spamming one template stops paying.

4 Experimental Setup

We post-train Qwen2.5-1.5B/3B-Instruct [7] from an SFT format-primed base (B0; trained on Boxoban `medium/train`). RL uses GRPO via the TRL library [6] for 500 steps, group size 8, lr $= 10^{-5}$, 4–8 inner iterations (so the rollout/policy ratio is non-trivial), bf16, no advantage standardization. We run 3 seeds per cell and log a diversity trajectory every 20 steps. B0 reaches $\sim$88% (1.5B) / 85% (3B) spec-valid syntax with full diversity (distinct = 1.0); a random-generator+solver-filter baseline (B1) also has distinct = 1.0 (the ceiling).

5 Results

5.1 A sharp reward↔diversity phase transition

Across all 3 objectives × 3 seeds × 2 scales, driving the gated reward to its maximum collapses diversity: *distinct@valid* falls from $\approx$1.0 to $<$0.05, typically within 50–150 steps, as in-band reward rises toward 1.0. Final-state diversity is near zero for essentially every high-reward run.

5.2 The trade-off is a controllable Pareto frontier

Plotting diversity against reward along training traces a clean frontier. At 1.5B one can hold distinct ≥ 0.5 up to reward ≈ 0.47, or distinct ≈ 0.40 at reward ≈ 0.69; the analogous 3B knee is reward 0.61 at distinct 0.70. **Early-stopping at the knee** is thus a simple passive control: pick the operating point on the frontier.

5.3 Active control: a diversity-aware reward (a small effect)

Reward (1.5B, vanilla, 3 seeds)	distinct@valid	distinct @ rew $\approx$0.8	peak	**cell-disagree**
Gated (standard)	0.004–0.016	0.019	0.199	**0.00**
Diversity-aware (ours)	0.051–0.095	0.155	0.371	**0.03**
un-collapsed (B0/start), ref.	1.0	n/a	n/a	*0.39*

Table 1: The diversity-aware reward induces *some* diversity at high reward (reward 0.77–1.0), raising the *distinct* count $\sim 8\times$. But the *structural* measure (mean fraction of differing cells between samples) moves only $0.00 \rightarrow 0.03$, far below the 0.39 of an un-collapsed model: it breaks exact duplication without restoring real variety. We caution that exact-match diversity metrics (distinct/entropy) overstate this effect.

Figure 1 shows why: among 32 sampled levels the gated model produces a single template; the diversity-aware model's "8 distinct" levels are mostly *one-cell* perturbations of one dominant template (21/32 identical), so they look nearly the same. The effect is real but small at 1.5B/3B; we hypothesize that larger-scale or longer training (which we show improves the diversity/reward frontier overall) is needed to promote it into *structural* diversity. Establishing that is left to future work.

5.4 Scale improves the frontier

The 3B diversity/reward Pareto frontier *dominates* the 1.5B frontier (more diversity at matched reward; e.g. reward $\approx$ 0.6 yields distinct 0.70 at 3B vs. $\approx$ 0.42 at 1.5B), even though the per-

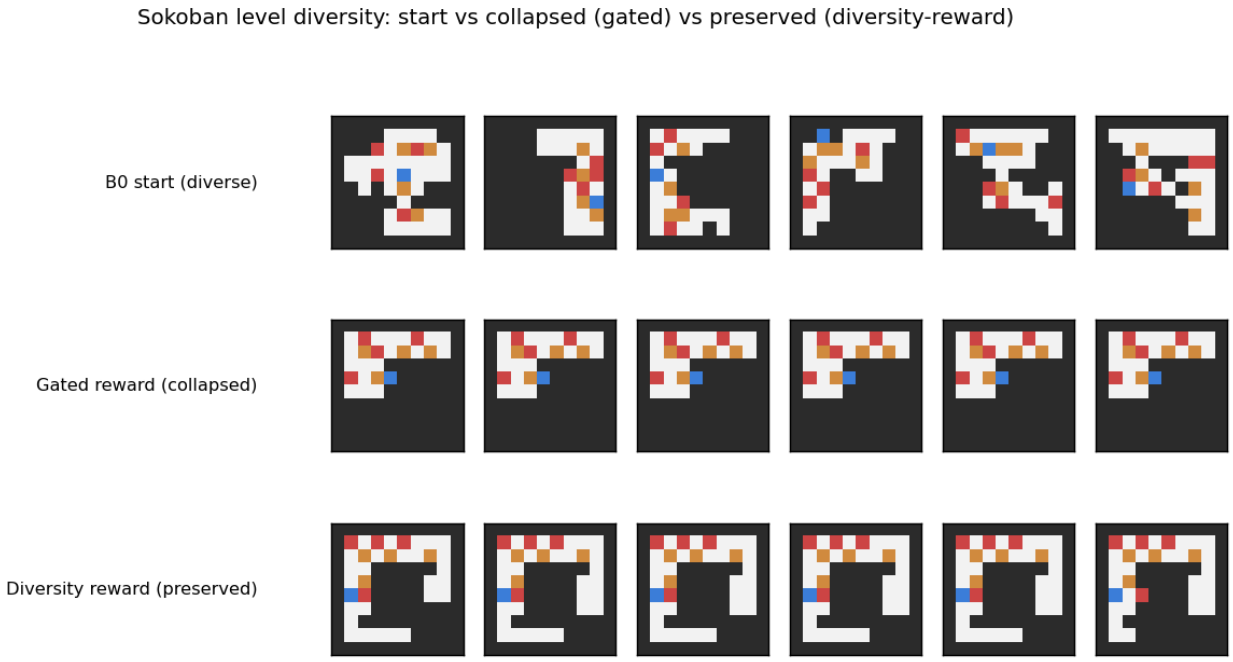

Figure 1: Generated Sokoban levels (#=wall, box=orange, goal=red, player=blue). Top: diverse start (32/32 distinct, cell-disagreement 0.39). Middle: gated reward collapses to one template (1/32, 0.00). Bottom: diversity-aware reward only modestly increases variety (its "8/32 distinct" are mostly one-cell perturbations of one template; cell-disagreement 0.03).

objective collapse dynamics are similar.

5.5 The trust-region objective is not a robust diversity lever

At 1.5B, DPPO retained the most diversity at matched high reward (distinct@reward0.8: DPPO 0.113, vanilla 0.064, clip-higher 0.022) and the highest peak (0.359 vs. 0.265), weakly supporting the hypothesis. However this **did not replicate at 3B** (DPPO had the *lowest* peak; the matched-reward ordering was within seed noise), and an initial single-seed result suggesting "vanilla collapses least" was overturned by seeding. We therefore conclude the objective choice has at most a small, inconsistent, scale-dependent effect on collapse dynamics, with no robust winner, a caution against assuming the clipping mechanism governs generative diversity on verifiable tasks.

6 Limitations

Single task (Sokoban, 10×10/4-box), two scales (1.5B, 3B), 3 seeds, modest step budget; one diversity-reward formulation and one DPPO δ. The 3B matched-reward comparison is partly limited by how fast collapse occurs relative to eval cadence. The diversity-reward effect, while large, has one of three seeds weaker. These bound the strength of the negative result and motivate larger-scale and broader-domain replication.

7 Conclusion

Verifiable-reward RL trades generative diversity for reward through a sharp phase transition; the trade-off is real, measurable, and, importantly, controllable. The reliable lever is operating-point selection (early-stopping at the Pareto knee), where genuine structural diversity still exists; an explicit diversity-aware reward induces only a small additional effect at our scale and is a candidate for larger-scale study. Larger models offer a better frontier overall, and the specific trust-region objective matters far less than commonly assumed.

Availability. Code, the SFT-primed checkpoints, the exact solver/verifier, and a reproduction recipe are available upon request.

References

[1] Arthur Guez, Mehdi Mirza, Karol Gregor, Rishabh Kabra, Sébastien Racanière, Théophane Weber, David Raposo, Adam Santoro, Laurent Orseau, Tom Eccles, Greg Wayne, David Silver, Timothy Lillicrap, and Victor Valdes. Boxoban Levels. `https://github.com/deepmind/boxoban-levels`, 2018.

[2] Ahmed Khalifa, Philip Bontrager, Sam Earle, and Julian Togelius. PCGRL: Procedural Content Generation via Reinforcement Learning. arXiv:2001.09212 [cs.LG], 2020. URL `https://arxiv.org/abs/2001.09212`.

[3] Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the Trust Region in LLM Reinforcement Learning. arXiv:2602.04879 [cs.LG], 2026. URL `https://arxiv.org/abs/2602.04879`.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG], 2017. URL `https://arxiv.org/abs/1707.06347`.

[5] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL], 2024. URL `https://arxiv.org/abs/2402.03300`.

[6] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformer Reinforcement Learning. `https://github.com/huggingface/trl`, 2020.

[7] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL], 2024. URL `https://arxiv.org/abs/2412.15115`.

[8] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476 [cs.LG], 2025. URL `https://arxiv.org/abs/2503.14476`.

[9] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization. arXiv:2507.18071 [cs.LG], 2025. URL `https://arxiv.org/abs/2507.18071`.