Entropy Regularizing Activation: Boosting Continuous Control, Large Language Models, and Image Classification with Activation as Entropy Constraints

Zilin Kang[1,2*], Chonghua Liao[3*], Tingqiang Xu[3*], Huazhe Xu[1,3,4]

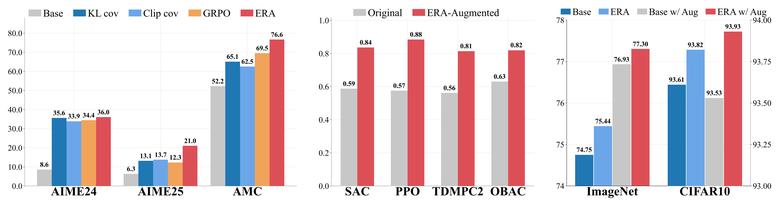
1 Shanghai Qi Zhi Institute; 2 Department of Computer Science and Technology, Tsinghua University; 3 Institute for Interdisciplinary Information Sciences, Tsinghua University; 4 Shanghai Artificial Intelligence Laboratory

The Entropy Dilemma

Encouraging exploration via policy entropy is critical in RL and LLM alignment.

The Problem: Standard methods (SAC, PPO, GRPO) add an entropy bonus directly to the loss function.

- Alters the optimization landscape.
- Causes **gradient conflicts** between reward and entropy.
- Leads to entropy collapse in LLMs.



(a) Large Language Models (b) Continuous Control (c) Image Classification

Figure 1. **ERA Boosts Large Language Models, Continuous Control and Image Classification.** (a) **Large Language Models:** ERA consistently enhances the performance of Qwen-2.5-Math-7B on AIME'24, AIME'25 and AMC datasets. (b) **Continuous Control:** ERA significantly improves multiple popular RL algorithms, including SAC, PPO, TD-MPC2 and OBAC. (c) **Image Classification:** ERA consistently boosts the performance of ResNet-50 on ImageNet and CIFAR-10 datasets.

Our Solution: ERA

Entropy Regularizing Activation (ERA) enforces maximum entropy via the output activation layer, not the loss function!

$$\pi_\theta(\cdot|s) = \pi_{g(f_\theta(s))}(\cdot|s) \quad \text{s.t.} \quad \mathbb{E}_{s\sim\rho_\pi}[\mathcal{H}_{\pi_\theta(\cdot|s)}] \geq \mathcal{H}_0$$

This approach completely decouples the optimization of the primary objective from the entropy constraint, allowing the loss function to focus solely on its original goal (e.g., maximizing rewards).

- **Decoupled:** Eliminates gradient conflicts.
- **Non-invasive:** Just swap the final network output layer.
- **Provable:** Guarantees a strict minimum sampling entropy constraint.

Continuous & Discrete Control

Continuous Control (Gaussian): ERA guarantees entropy by adjusting the standard deviation $\hat{\sigma}$ via a dimension-aware softmax. It learns to allocate entropy optimally across action dimensions, unlike baseline methods that scale uniformly.

$$\sigma_i' = \exp\left[\max\left(\log\sigma_{\max} + \left(\mathcal{H}_0' - D\log\sqrt{2\pi e} - D\log\sigma_{\max}\right)\frac{e^{\hat{\sigma}_i}}{\sum_{j=1}^{D} e^{\hat{\sigma}_j}}, \log\sigma_{\min}\right)\right]$$

Discrete Classification (Softmax): Acts as a dynamic, sample-specific label smoothing that prevents overconfidence.

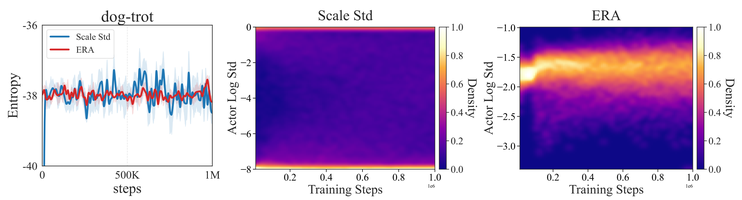


Figure 2. Evolution of log standard deviation on Dog-trot. ERA learns a balanced, diffusive allocation compared to polarized standard scaling baselines.

Adaptive ERA for LLMs

Forcing high entropy on *all* language tokens causes gibberish. We introduce an **adaptive ERA** applied during the GRPO update step. We monitor H_{resp} (**average entropy of the top 20% forking tokens**), and dynamically transform the logits ($z \to z'$) and advantages ($A_t \to A_t'$):

$$z' = \begin{cases} kz & \text{if } H_{\text{resp}} < \omega_{\text{low}}, A_t > 0 \\ \frac{1}{k}z & \text{if } H_{\text{resp}} > \omega_{\text{high}}, A_t > 0 \\ z & \text{otherwise} \end{cases} \qquad A_t' = \begin{cases} \frac{1}{k}A_t & \text{if } H_{\text{resp}} < \omega_{\text{low}}, A_t > 0 \\ kA_t & \text{if } H_{\text{resp}} > \omega_{\text{high}}, A_t > 0 \\ A_t & \text{otherwise} \end{cases}$$

- **Sharpening ($H_{\text{resp}} < \omega_{\text{low}}$):** Increases probabilities for positively advantaged responses to boost exploration.
- **Flattening ($H_{\text{resp}} > \omega_{\text{high}}$):** Depresses probabilities to prevent overly random text.

This formulation acts as an adaptive KL regularizer, strictly preventing entropy collapse in on-policy LLM training.

Breakthroughs in Continuous Control

ERA accelerates learning across multiple algorithms (**SAC, PPO, FastSAC, TD-MPC2**).

On the hardest **HumanoidBench** and **Dog** suites, SAC-ERA improves performance by >25% over standard SAC.

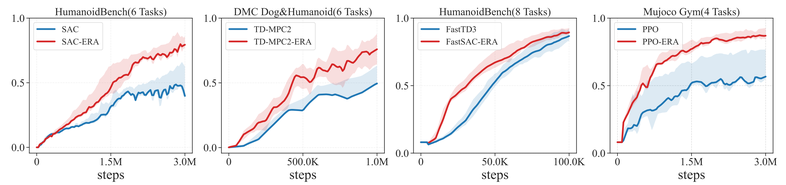


Figure 3. **Normalized Performance:** ERA dominates asymptotic performance and sample efficiency.

Massive LLM Improvements

Applied to **Qwen2.5-Math-7B** via GRPO, ERA establishes a non-trivial entropy floor and preserves diverse reasoning trajectories.

Table 1. **Scores on Math Reasoning Benchmarks (%)**

Method	AIME24	AIME25	AMC	MATH500	Minerva	Olympiad	Avg.
GRPO	34.4	12.3	69.5	80.6	36.8	40.6	45.7
ERA (Ours)	**36.0**	**21.0**	**76.6**	**85.4**	**40.1**	**46.8**	**51.0**
Δ (Relative)	+4.7%	+70.7%	+10.4%	+6.0%	+9.0%	+15.3%	+11.6%

Pass@k results confirm that maintaining a healthy entropy floor directly translates to better mathematical reasoning.

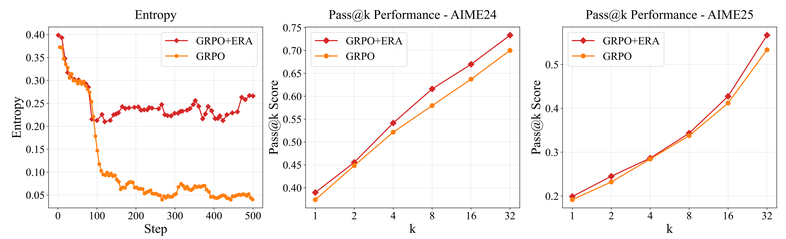


Figure 4. **GRPO vs. GRPO+ERA:** ERA completely stabilizes entropy (left), drastically increasing exploration and Pass@k success (right).