SAMPPO: Structure-Aware Mirror Proximal Policy Optimization

Corinna Cortes
Google Research
New York, NY 10011
corinna@google.com

Mehryar Mohri
Google Research & CIMS
New York, NY 10011
mohri@google.com

Yutao Zhong
Google Research
New York, NY 10011
yutaozhong@google.com

Abstract

The alignment of Large Language Models (LLMs) via RLHF typically relies on Proximal Policy Optimization (PPO) with a uniform clipping parameter ϵ to maintain a stable trust region. We show that this uniform constraint is ill-suited for the non-uniform semantic topology of natural language: in *semantically dense* regions, where candidate tokens act as contextual synonyms, noisy advantage estimates destabilize the trust region, causing *policy collapse*: a significant loss of generative diversity and entropy. We propose *Structure-Aware Mirror Proximal Policy Optimization (SAMPPO)*, which introduces an adaptive trust region governed by local semantic density. Using a per-timestep *Semantic Isolation Score*, SAMPPO dynamically tightens clipping in dense regions and relaxes it in sparse ones. We derive a *Structure-Aware Monotonic Improvement* bound, establish a formal link to margin-shifted $\mathcal{H}$-consistency, and prove that the adaptive schedule realizes a uniform weighted-KL budget, which we verify empirically. On HH-RLHF and HelpSteer with Llama-3.1-8B and Gemma-2-2B, SAMPPO prevents entropy collapse, consistently outperforms PPO and BAPO across LLM-judge, reward-model, and diversity evaluations, and remains competitive with DPO and SimPO at $+1.9\%$ wall-clock overhead.

1 Introduction

The alignment of Large Language Models (LLMs) with human intent has largely transitioned from explicit reward modeling and traditional reinforcement learning [Christiano et al., 2017, Stiennon et al., 2020, Ouyang et al., 2022] to direct policy optimization. Methods such as DPO [Rafailov et al., 2023], IPO [Azar et al., 2024], KTO [Ethayarajh et al., 2024], and SimPO [Meng et al., 2024] have emerged as popular standards, allowing practitioners to fine-tune models directly on pairwise preference data without the instability of Proximal Policy Optimization (PPO) [Schulman et al., 2017]. However, a key theoretical deficiency exists in this paradigm: for the equicontinuous hypothesis sets typical of neural networks, standard surrogate losses, such as the logistic loss used in DPO, are theoretically inconsistent [Mohri and Zhong, 2026a].

This inconsistency is particularly pronounced in *semantically dense* action spaces where the model must distinguish between highly similar responses, such as candidate tokens acting as semantic synonyms within the current context. Standard PPO, widely used in foundational RLHF [Bai et al., 2022], restricts updates using a uniform clipping hyperparameter ϵ, while DPO implicitly enforces a uniform trust region via a fixed KL-divergence penalty β: both treat all actions as equally distinct. This one-size-fits-all approach fails when noisy advantage estimates destabilize the fixed trust region, causing the policy to over-optimize an arbitrary token while suppressing its viable semantic neighbors. This phenomenon, which we refer to as *policy collapse*, degrades linguistic diversity and generative quality to overfit preference noise, a vulnerability recently observed empirically in reasoning models as well [Cui et al., 2025].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We introduce *Structure-Aware Mirror Proximal Policy Optimization (SAMPPO)* to address this instability by transitioning from a fixed, uniform trust region to an *adaptive trust region* that responds to the local semantic topology of the action space. Integrating a computable distance pre-metric $\Delta(a, a')$ between actions, SAMPPO evaluates a per-timestep *Semantic Isolation Score* $D_{\pi_{\text{old}}}(s_t, a_t)$ that dynamically tightens the clipping range ϵ_t in dense regions, preventing the exploitation of noisy signals among synonyms, while relaxing it in sparse regions to facilitate efficient learning.

Related Work. Our work builds on direct preference alignment [Rafailov et al., 2023, Azar et al., 2024, Meng et al., 2024], adaptive trust-region methods for PPO [Wang et al., 2019, Gao et al., 2025], and recent $\mathcal{H}$-consistency theory for surrogate losses in preference learning [Mohri and Zhong, 2026a]. In particular, Xi et al. [2025] observe a similar empirical "entropy collapse" phenomenon and propose BAPO, an adaptive clipping strategy; however, BAPO relies on an empirical search heuristic and lacks formal convergence guarantees, whereas SAMPPO derives its adaptive bounds directly from the geometric semantic topology of the action space. A detailed discussion of related work, including recent RLHF scaling systems [Yu et al., 2025] and entropy adaptations [Rahman, 2025, Cui et al., 2025], is provided in Appendix A.

Main Contributions. In summary, this work makes four key contributions:

- *Formalization of Semantic Density:* We formalize the semantic density of actions and prove that uniform trust regions are vulnerable to entropy collapse even under exchangeable noise (Corollary 4).
- *The SAMPPO Algorithm:* We propose an algorithm with a state-action-dependent adaptive clipping mechanism using local semantic structure to stabilize policy optimization.
- *Structure-Aware Consistency Bounds:* We derive monotonic improvement bounds preserving trust-region stability [Schulman et al., 2015] in non-uniform spaces, prove a uniform global weighted-KL budget (Proposition 8), and establish a link to $\mathcal{H}$-consistency.
- *Empirical Validation:* Using Llama-3.1-8B [Llama Team, AI @ Meta, 2024] and Gemma-2-2B [Gemma Team, Google DeepMind, 2024] on HH-RLHF [Bai et al., 2022] and HelpSteer [Wang et al., 2024b], SAMPPO outperforms PPO and BAPO [Xi et al., 2025] across LLM-judge, reward-model, and diversity metrics at a measured $+1.9\%$ wall-clock overhead, with cross-dataset hyperparameter transferability, while remaining highly competitive with DPO and SimPO. A validation study confirms that the Semantic Isolation Score captures contextual synonym density rather than policy confidence, token frequency, or anisotropy.

2 The Problem: Failure of the Uniform Trust Region

2.1 Action Space Geometry and Semantic Isolation

We begin by characterizing the structure of the action space $\mathcal{A}$; in autoregressive language modeling, an action $a \in \mathcal{A}$ is the selection of the *next candidate token* from the vocabulary.

Definition 1 (Semantic Action Distance Pre-Metric)**.** *Let $\Delta: \mathcal{A} \times \mathcal{A} \to [0, \Delta_{\max}]$ be a bounded, symmetric computable pre-metric on the action space. In the context of language models, Δ is typically defined as the cosine distance between the candidate token embeddings: $\Delta(a, a') = 1 - \cos(E(a), E(a'))$, where $E(a)$ denotes the embedding vector of token a, either its input embedding (used in our main experiments) or its output unembedding (`lm_head`) vector (compared in Section 5.2; see also Appendix B). Note, the mapped point $E(a)$ in the embedding space evaluates the semantic representation of the individual candidate token itself, independent of the preceding sequence context.*

Using continuous semantic space metrics to evaluate structural geometry is supported by recent uncertainty quantification literature [Qiu and Miikkulainen, 2024]. A pre-metric may not verify the triangle inequality, but this property is not required in our proofs or algorithm.

Definition 2 (Semantic Isolation Score)**.** *For a state s and a fixed reference policy π_{old}, the Semantic Isolation Score of an action (candidate token) a is defined as:*

$$D_{\pi_{\text{old}}}(s, a) = \mathop{\mathbb{E}}_{a' \sim \pi_{\text{old}}(\cdot|s)} [\Delta(a, a')]. \tag{1}$$

Taking the expectation over $\pi_{\text{old}}(\cdot|s)$ makes $D_{\pi_{\text{old}}}(s, a)$ context-dependent: a candidate token a is *semantically dense* if $D_{\pi_{\text{old}}}(s, a) \to 0$, i.e., the policy's probability mass is concentrated on a cluster of near-synonyms in the current context (additional comments in Appendix B).

2.2 Entropy Collapse in Semantically Dense Regions

We now formalize the phenomenon by establishing a quantitative entropy-collapse result; a similar empirical concern during LLM fine-tuning was recently raised by Xi et al. [2025]. In semantically dense regions, even a single noisy advantage estimate induces a systematic reduction in policy entropy under clipped updates.

Setup. Consider a fixed state s and a cluster of $\kappa \geq 2$ semantically indistinguishable actions (candidate tokens) $\{a_1, \ldots, a_\kappa\}$ with $\Delta(a_i, a_j) \leq \epsilon_{\text{sem}}$ for a small threshold ϵ_{sem}. Assume the reference policy is uniform over this set, $\pi_{\text{old}}(a_i \mid s) = \frac{1}{\kappa}$, and the reward model assigns a noisy reward $R > 0$ to a_1 and 0 to all synonyms a_j, $j \geq 2$. As is standard in PPO via Generalized Advantage Estimation (GAE), advantages are mean-centered; with mean reward $\overline{R} = R/\kappa$, the centered advantages are $\widehat{A}(s, a_1) = \frac{\kappa-1}{\kappa} R > 0$ and $\widehat{A}(s, a_j) = -\frac{1}{\kappa} R < 0$ for $j \geq 2$.

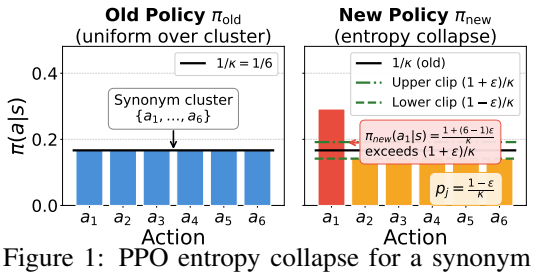


Figure 1: PPO entropy collapse for a synonym cluster of size $\kappa = 6$ and $\epsilon = 0.15$.

Since *all* synonyms a_j carry strictly negative advantages, the PPO surrogate gradient pushes every ratio $r_j = \pi_{\text{new}}(a_j|s)/\pi_{\text{old}}(a_j|s)$ with $j \geq 2$ downward until they saturate the lower clip bound at $1-\epsilon$, so $\pi_{\text{new}}(a_j|s) = (1-\epsilon)/\kappa$ for all $j \geq 2$. The simplex constraint $\sum_j \pi_{\text{new}}(a_j|s) = 1$ then forces a_1 to absorb all displaced mass:

$$\pi_{\text{new}}(a_1|s) = \frac{1 + (\kappa - 1)\epsilon}{\kappa}, \qquad \pi_{\text{new}}(a_j|s) = \frac{1-\epsilon}{\kappa} \quad \text{for } j \geq 2. \tag{2}$$

This is the global optimum of the clipped surrogate on the simplex: r_j is clipped at $1-\epsilon$, while $r_1 = 1 + (\kappa - 1)\epsilon > 1 + \epsilon$ for all $\kappa > 2$ falls in the flat region of the upper clip, so no gradient opposes the simplex-forced redistribution: the simplex constraint, not the gradient, drives $\pi_{\text{new}}(a_1|s)$ *beyond* the upper clip bound. Figure 1 illustrates the case $\kappa = 6$, $R = 1$, $\epsilon = 0.15$. In practical PPO, finite mini-batch gradient steps approach this stationary point asymptotically, and repeated iterations compound the effect.

Theorem 3 (Entropy Collapse under Mean-Centered Noisy Advantage)**.** *Under the described setup, assuming a small-cluster regime with $(\kappa - 1)\epsilon < 1$, the PPO update produces a policy with entropy:*

$$H(\pi_{new}(\cdot \mid s)) = \log \kappa - \frac{(\kappa-1)\epsilon^2}{2} + \frac{(\kappa-1)(\kappa-2)\epsilon^3}{6} + \mathcal{O}(\epsilon^4). \tag{3}$$

Thus, the entropy drop is strictly positive for $\epsilon > 0$, approximated by: $H(\pi_{old}(\cdot \mid s)) - H(\pi_{new}(\cdot \mid s)) \approx \frac{\kappa-1}{2}\epsilon^2$. For larger clusters outside the convergence radius of this expansion ($\kappa\epsilon \gg 1$), the exact entropy drop scales logarithmically as $\mathcal{O}(\epsilon \log \kappa)$. In both regimes, the entropy drop scales monotonically *with cluster size κ: semantically denser regions suffer greater entropy collapse.*

Proof Sketch. The PPO update drives unrewarded synonyms to the lower clip bound, and the sum-to-one constraint forces the rewarded synonym's probability upward. A third-order expansion of the new entropy cancels the linear terms, leaving a positive drop dominated by $\frac{\kappa-1}{2}\epsilon^2$ (Appendix E.1). □

Theorem 3 analyzes a specific noise realization in which a_1 draws the positive reward. Could an unbiased, *exchangeable* noise model, with the favored synonym varying across realizations and a symmetric expected update, avoid systematic collapse? It cannot:

Corollary 4 (Expected Entropy Decay under Exchangeable Noise)**.** *Under the setup of Theorem 3, suppose the rewarded synonym is drawn uniformly at random from the cluster (exchangeable noise), so that $\mathbb{E}[\pi_{new}(\cdot|s)] = \pi_{old}(\cdot|s)$. Then, by symmetry, the one-step entropy drop is the same deterministic quantity for every realization: $\mathbb{E}[H(\pi_{new})] = H(\pi_{old}) - \frac{\kappa-1}{2}\epsilon^2 + \mathcal{O}(\epsilon^3)$. More generally, for* any *unbiased stochastic update with $\mathbb{E}[\pi_{new}] = \pi_{old}$ uniform on the cluster, Jensen's inequality applied to the strictly concave entropy functional yields $\mathbb{E}[H(\pi_{new})] \leq H(\mathbb{E}[\pi_{new}]) = H(\pi_{old})$, with strict inequality unless the update is degenerate.*

Thus, continuous zero-mean stochastic oscillation systematically erodes entropy: PPO's non-linear clipping, coupled with the simplex constraint, acts as a one-way mechanism that symmetry of the expected update cannot undo (proof in Appendix E.2).

3 The SAMPPO Algorithm

To address the entropy decay established above, we introduce Structure-Aware Mirror Proximal Policy Optimization (SAMPPO) with an adaptive mechanism that respects the semantic topology of the action space.

Algorithm 1 Structure-Aware Mirror PPO (SAMPPO)

Initialize policy θ, value function ϕ, and hyperparameters $\epsilon_{\min}, \epsilon_{\max}, k, D_{\text{bound}}$
for iteration = 1, 2, … **do**
  Collect trajectories $\mathcal{D}_{\text{iter}}$ using behavior policy π_{old}
  Compute advantages $\widehat{A}_t$ via GAE
  //— SAMPPO Static Preparation —
  for each timestep t in $\mathcal{D}_{\text{iter}}$ **do**
    Extract top-k actions $\{a'_i\}_{i=1}^{k}$ from $\pi_{\text{old}}(\cdot|s_t)$
    Normalize probabilities $p_i = \pi_{\text{old}}(a'_i|s_t) / \sum_{j=1}^{k} \pi_{\text{old}}(a'_j|s_t)$
    Compute deterministic score $D_{\pi_{\text{old}}}(s_t, a_t) \approx \sum_{i=1}^{k} p_i \Delta(a_t, a'_i)$
    Compute scaling $S_t = \min(D_{\pi_{\text{old}}}(s_t, a_t)/D_{\text{bound}}, 1.0)$
    Set adaptive $\epsilon_t = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot S_t$ **// Apply** `stop_gradient`
  end for
  //— Surrogate Optimization —
  for epoch = 1, 2, …, N **do**
    Update θ by maximizing $L^{\text{SAMPPO}}(\theta)$ via gradient ascent over $\mathcal{D}_{\text{iter}}$, treating ϵ_t as a constant
  end for
  Update ϕ and set $\pi_{\text{old}} \leftarrow \pi_\theta$
end for

Objective Function: SAMPPO incorporates the Semantic Isolation Score D into an adaptive ϵ_t that keeps the trust region conservative in semantically dense regions and expansive in sparse ones. Given the probability ratio $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$ and the advantage estimate $\widehat{A}_t$, the modified objective is defined by:

$$L^{\text{SAMPPO}}(\theta) = \hat{\mathbb{E}}_t\big[\min\big(r_t(\theta)\widehat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon_t, 1+\epsilon_t)\widehat{A}_t\big)\big], \tag{4}$$

where θ denotes the policy model parameters and $\epsilon_t = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot S_t$, with $S_t = \min(D_{\pi_{\text{old}}}(s_t, a_t)/D_{\text{bound}}, 1.0)$. Here, D_{bound} reflects the geometry of the LLM's highly anisotropic representation space [Ethayarajh, 2019] and is easily estimated experimentally (see Appendix G for the geometric justification); Appendix C provides additional detail on the algorithm and its properties.

Pseudocode: SAMPPO (see Algorithm 1) requires minimal overhead beyond standard PPO. The Semantic Isolation Score is computed once per rollout step outside the inner epoch loop from only the top-k tokens of the frozen behavior policy, avoiding both Monte Carlo variance and full-vocabulary cost. Since ϵ_t is detached from the computation graph, no gradients flow through the clipping bounds, adding negligible overhead and requiring no differentiability of the distance metric.

4 Theoretical Foundations: Structure-Aware Improvement

4.1 Structure-Aware Divergence and Performance Bound

In standard trust-region methods, policy updates are controlled via a global divergence such as the Kullback–Leibler (KL) divergence, which ignores the varying semantic density of structured action spaces. To address this, we define a structure-aware expected weighted divergence.

Definition 5 (Structure-Aware Weighted KL Divergence)**.** *Let $D_{\pi_{old}}(s, a)$ denote the Semantic Isolation Score and define the expected state-level isolation: $D(s) = \mathbb{E}_{a\sim\pi_{old}(\cdot|s)}[D_{\pi_{old}}(s, a)]$. For a*

fixed stability constant $\nu > 0$, define the weight: $w(s) = \frac{1}{D(s)+\nu}$. Let $\rho_{\pi_{old}}$ denote the state visitation distribution induced by π_{old}. The structure-aware divergence between policies is defined via the KL divergence as:

$$D_w(\pi, \pi_{old}) = \mathop{\mathbb{E}}_{s\sim\rho_{\pi_{old}}}\Big[w(s)\mathsf{D}_{\mathrm{KL}}\big(\pi(\cdot|s) \parallel \pi_{old}(\cdot|s)\big)\Big]. \tag{5}$$

The additive constant ν keeps the weights bounded in regions of extremely high semantic density. The following lemma establishes that this divergence is equivalent, up to bounded constants, to the standard average KL divergence, ensuring it is a well-behaved metric for optimization.

Lemma 6 (Equivalence to Average KL)**.** *Let the state-level isolation score $D(s)$ be bounded such that $0 \le D_{\min} \le D(s) \le D_{\max}$. Let the weight function be $w(s) = \frac{1}{D(s)+\nu}$ for some constant $\nu > 0$. The structure-aware divergence $D_w(\pi, \pi_{old})$ and the standard average KL divergence $\mathsf{D}_{\mathrm{KL}}^{avg}$ satisfy the following two-sided inequality:*

$$(D_{\min} + \nu) \cdot D_w(\pi, \pi_{old}) \le \mathsf{D}_{\mathrm{KL}}^{avg} \le (D_{\max} + \nu) \cdot D_w(\pi, \pi_{old}). \tag{6}$$

Proof Sketch. Substituting the bounds $D_{\min}$ and $D_{\max}$ into the relationship between $w(s)$ and the average KL divergence gives a direct pointwise bound on the expectation (Appendix E.3). □

We now establish a monotonic improvement bound using $D_w(\pi, \pi_{\text{old}})$, demonstrating how the adaptive trust region preserves performance and prevents entropy collapse.

Theorem 7 (Structure-Aware Surrogate Improvement)**.** *Let $C_0 = \frac{2\gamma}{(1-\gamma)^2}\max_{s,a}|A_{\pi_{old}}(s,a)|$ and $C = C_0 \cdot (D_{\max} + \nu)$, where $J(\pi)$ denotes the expected return.*

(a) ***Rigorous bound.*** *The structure-aware surrogate objective satisfies:*

$$J(\pi) \ge J(\pi_{old}) + L_{\pi_{old}}(\pi) - C \cdot \max_s\Big[w(s)\mathsf{D}_{\mathrm{KL}}\big(\pi(\cdot|s) \parallel \pi_{old}(\cdot|s)\big)\Big]. \tag{7}$$

(b) ***Practical surrogate.*** *As in Schulman et al. [2015], replacing the intractable worst-case maximum with the expected value under $\rho_{\pi_{old}}$ yields the surrogate used in practice:*

$$\widetilde{J}(\pi) = J(\pi_{old}) + L_{\pi_{old}}(\pi) - C \cdot D_w(\pi, \pi_{old}). \tag{8}$$

Since $\mathbb{E}_s[\cdot] \le \max_s[\cdot]$, the surrogate $\widetilde{J}(\pi)$ upper-bounds the rigorous bound (7); it is therefore a less conservative (but still consistent) proxy for policy improvement.

(c) ***Rigorous average-divergence bound.*** *Let $\zeta^\pi = \max_s\big|\mathbb{E}_{a\sim\pi(\cdot|s)}[A_{\pi_{old}}(s,a)]\big|$, so that $\frac{2\gamma\zeta^\pi}{(1-\gamma)^2} \le C_0$. Then*

$$J(\pi) \ge J(\pi_{old}) + L_{\pi_{old}}(\pi) - \frac{2\gamma\zeta^\pi}{(1-\gamma)^2}\sqrt{\frac{D_{\max}+\nu}{2}\,D_w(\pi, \pi_{old})}. \tag{9}$$

Proof Sketch. The standard trust-region bound controls the performance difference by the maximum KL divergence across states; bounding the pointwise KL by $(D_{\max} + \nu)\, w(s)\, \mathsf{D}_{\mathrm{KL}}$ yields part (a). Replacing the intractable worst-case maximum by the tractable expectation D_w, as in Schulman et al. [2015], gives the practical surrogate in part (b). Part (c) follows from the performance-difference bound of Achiam et al. [2017], which holds for the total variation distance *averaged* over the state distribution, combined with Pinsker's and Jensen's inequalities (Appendix E.4). □

Part (c) gives a rigorous grounding to the average-case penalty: whereas part (b) is obtained by the standard practical relaxation of Schulman et al. [2015], the bound of Achiam et al. [2017] holds directly for the state-averaged divergence, so the average structure-aware divergence D_w *provably* controls the performance difference, at the cost of a square-root dependence and with a leading constant no larger than C_0.

4.2 From Adaptive Clipping to a Uniform Structure-Aware KL Budget

The adaptive parameter ϵ_t induces a localized trust region: $\epsilon_t \to \epsilon_{\min}$ in dense regions, preventing exploitation of noisy signals among indistinguishable actions, and $\epsilon_t \to \epsilon_{\max}$ in sparse regions. We now make this precise. Clipped surrogates do not hard-enforce a divergence constraint, just as PPO

does not hard-enforce TRPO's KL trust region, and Theorem 3 shows the simplex constraint can push a realized ratio past the clip. What *can* be established is that, under the idealized clip condition the surrogate is constructed to approximate, the schedule ϵ_t provably realizes the structure-aware weighted-KL constraint with a *single uniform global budget*.

Proposition 8 (Adaptive Clipping Realizes a Uniform Structure-Aware KL Budget)**.** *For each state s, let $\epsilon(s) = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot \min(D(s)/D_{bound}, 1)$ denote the clip radius evaluated at the state-level isolation score $D(s)$, and suppose the updated policy satisfies the idealized clip condition $|\pi(a|s)/\pi_{old}(a|s) - 1| \le \epsilon(s)$ for all a, and assume $\epsilon_{\max} < 1$. Then*

$$\mathsf{D}_{\mathrm{KL}}(\pi(\cdot|s) \,\|\, \pi_{old}(\cdot|s)) \le \frac{\epsilon(s)^2}{2(1-\epsilon_{\max})},$$

and therefore, with $w(s) = 1/(D(s)+\nu)$:

$$w(s)\mathsf{D}_{\mathrm{KL}}(\pi(\cdot|s) \,\|\, \pi_{old}(\cdot|s)) \le \delta^* := \sup_s \frac{w(s)\epsilon(s)^2}{2(1-\epsilon_{\max})} \le \frac{\epsilon_{\min}^2/\nu + (\epsilon_{\max}^2 - \epsilon_{\min}^2)/D_{bound}}{2(1-\epsilon_{\max})} < +\infty,$$

uniformly over all states.

Proof Sketch. Writing the per-state KL as $\sum_a \pi_{\text{old}}(a)\varphi(r(a))$ with $\varphi(x) = x\log x - x + 1$, Taylor's theorem with $\varphi(1) = \varphi'(1) = 0$ and $\varphi'' \le 1/(1-\epsilon_{\max})$ on the clip interval bounds each term by $\epsilon(s)^2/(2(1-\epsilon_{\max}))$. The uniform bound δ^* follows since $\epsilon(\cdot)$ is affine below D_{bound} and constant above it. The full proof is given in Appendix E.5. □

Relative to a uniform clip $\epsilon = 0.20$, the schedule tightens the weighted-KL budget in dense regions by a factor $(\epsilon/\epsilon_{\min})^2 \approx 4$ while permitting steps up to $\epsilon_{\max}$ in sparse ones, avoiding $\mathcal{O}(|\mathcal{V}|)$ full-vocabulary overhead. Appendix E.5 shows the per-action schedule is more conservative in expectation (Jensen) and that the penalized form recovers D_w exactly. *Empirical verification:* with $\nu = 0.05$ ($\delta^* = 0.33$), 99.6% of logged states satisfy $w(s)\mathsf{D}_{\mathrm{KL}} \le \delta^*$; rare exceedances (median +8%) are consistent with Theorem 3's simplex coupling (Appendix E.5.1).

4.3 Mirror Descent Interpretation and Convergence

Standard policy optimization can be viewed as mirror descent with a uniform KL divergence. SAMPPO replaces this with D_w, inducing a state-dependent Bregman divergence and the modified update $\pi_{k+1} = \operatorname{argmax}_\pi \left\{ \mathbb{E}_{s,a\sim\pi_k}\left[A_k(s,a)\frac{\pi(a|s)}{\pi_k(a|s)}\right] - \frac{1}{\eta}D_{w_k}(\pi,\pi_k)\right\}$. Under a local quadratic approximation, this is equivalent to natural policy gradient with a state-weighted Fisher information matrix $F_{w_k}(\theta_k) = \mathbb{E}_{s\sim\rho_{\pi_k}}[w_k(s)F_s(\theta_k)]$, which up-weights the local metric in dense regions (forcing smaller steps) and down-weights it in sparse regions: SAMPPO acts as *adaptive preconditioning* of the gradient according to the semantic structure of the action space (Appendix D).

To provide a global convergence guarantee beyond the one-step bound of Theorem 7, we establish that $D_{w_k}(\pi,\pi_k)$ is a valid Bregman divergence generated by the state-weighted negative entropy $\Psi_{w_k}(\pi) = \mathbb{E}_{s\sim\rho_{\pi_k}}[w_k(s)\sum_a \pi(a|s)\log\pi(a|s)]$ (Lemma 11, Appendix D.5), strongly convex with respect to the expected squared ℓ_1-norm with *uniform global modulus* $\mu = \frac{1}{D_{\max}+\nu} > 0$ (via Pinsker's inequality and $w_k(s) \ge \frac{1}{D_{\max}+\nu}$). Since this uniform lower bound holds across all iterations, the SAMPPO update admits a formal global convergence guarantee.

Theorem 9 (Convergence to a Stationary Point)**.** *Assume the expected policy objective $J(\pi)$ is L-smooth uniformly across shifting state distributions ρ_{π_k} with respect to the expected squared ℓ_1-norm ($\mathbb{E}_s[\|\pi(\cdot|s) - \pi_k(\cdot|s)\|_1^2]$), implicitly absorbing a worst-case concentrability coefficient standard in MDP analysis, and bounded above by J^*. Let the sequence of policies $\{\pi_k\}_{k=0}^T$ be generated by the SAMPPO mirror descent update:*

$$\pi_{k+1} = \operatorname*{argmax}_\pi \left\{ \langle \nabla J(\pi_k), \pi\rangle - \frac{1}{\eta}D_{w_k}(\pi,\pi_k)\right\}. \tag{10}$$

For a step size $\eta \le \frac{\mu}{L}$, SAMPPO converges to a stationary point at a rate of $\mathcal{O}(1/T)$:

$$\min_{0\le k<T} \mathbb{E}_s\left[\|\pi_{k+1} - \pi_k\|_1^2\right] \le \frac{2\eta(D_{\max}+\nu)}{T}(J^* - J(\pi_0)). \tag{11}$$

Proof Sketch. The first-order optimality condition of the mirror descent update bounds the improvement by the symmetrized Bregman divergence; uniform μ-strong convexity gives a strictly positive descent coefficient $\frac{\mu}{2\eta}$, and telescoping over T iterations yields the $\mathcal{O}(1/T)$ rate (Appendix E.7). □

This result confirms that SAMPPO is not merely a stabilizing heuristic but a principled optimization algorithm: in dense regions the large weight $w_k(s)$ scales up the local strong convexity of the penalty, forcing smaller, more conservative steps and ensuring stable global convergence.

4.4 Connection to Margin-Shifted $\mathcal{H}$-Consistency

We further ground SAMPPO's adaptive clipping mechanism in statistical learning theory. Recent advancements in $\mathcal{H}$-consistency for preference learning demonstrate that for restricted equicontinuous hypothesis sets (e.g., neural networks), standard unconstrained surrogate losses yield vacuous generalization bounds; consistency requires enforcing a separation margin γ via a *margin-shifted surrogate* $\Phi_\gamma(u) = \Phi(u - \gamma)$ (formal definitions in Appendix F).

In the offline setting, Mohri and Zhong [2026a] resolve this via Structure-Aware DPO with a dynamic pair-dependent margin $\Gamma(y, y') = \tau\Delta(y, y')$. We show that SAMPPO's adaptive trust region is the online counterpart: by bounding the surrogate update, it implicitly optimizes a structure-aware margin-shifted surrogate. The local equivalence relies on a first-order Taylor expansion whose approximation error is quantified in Proposition 14 (Appendix E.8).

Theorem 10 (Implicit Structure-Aware $\mathcal{H}$-Consistency)**.** *Let $h_\theta(s, a) = \log \pi_\theta(a|s) - \log \pi_{\text{old}}(a|s)$ be the relative scoring function of the policy hypothesis set $\mathcal{H}$. Up to a second-order approximation error uniformly bounded within the trust region (as explicitly quantified in Proposition 14), the SAMPPO clipped objective is mathematically equivalent to minimizing an advantage-weighted structure-aware margin-shifted Hinge surrogate:*

$$\mathsf{L}_{\Phi,\Gamma}(h_\theta) = |A_t|\Phi_{\text{hinge}}(y_t \cdot h_\theta(s_t, a_t) - \Gamma(s_t, a_t)), \tag{12}$$

where $y_t = \text{sign}(A_t)$, $\Phi_{\text{hinge}}(u) = \max(0, -u)$, and the local margin is $\Gamma(s_t, a_t) = \epsilon_t$. Thus, SAMPPO satisfies the established structure-aware $\mathcal{H}$-consistency bound:

$$\mathcal{R}(h_\theta) - \mathcal{R}^*(\mathcal{H}) + \mathcal{M}(\mathcal{H}) \le \mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(h_\theta) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H}) + \mathcal{A}_\Gamma(\mathcal{H}), \tag{13}$$

where $\mathcal{R}(h_\theta)$ is the target 0-1 ranking error, $\mathcal{M}(\mathcal{H})$ is the minimizability gap, $\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma} = \mathsf{L}_{\Phi_{\text{hinge}},\Gamma}/\Phi_{\text{hinge}}(-\Gamma)$ is the inverse-margin weighted surrogate loss, and $\mathcal{A}_\Gamma(\mathcal{H}) = \mathbb{E}[\mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H})] - \mathbb{E}[\mathcal{C}^*(\mathcal{H})]$ is the structure-aware approximation gap.

Proof Sketch. A first-order Taylor expansion of the probability ratio around the clipping bounds demonstrates that, for both positive and negative advantages, the optimization reduces to an advantage-weighted, margin-shifted Hinge surrogate. This equivalence holds strictly within the trust region up to the quantified bound formalized in Proposition 14. Since the local margin $\Gamma(s_t, a_t) = \epsilon_t$ scales proportionally to the local semantic isolation and its gradient is statically detached, it fulfills all structural requirements for $\mathcal{H}$-consistency. The detailed proof is in Appendix E.8. □

Comparative guarantee vs. uniform PPO. Whenever PPO's fixed $\epsilon = 0.20$ exceeds the model's semantic capacity $C_\mathcal{H}$ (which collapses toward 0 in dense synonym clusters), the approximation gap $\mathcal{A}_\Gamma(\mathcal{H})$ saturates, rendering the consistency bound vacuous (Proposition 16). By shrinking ϵ_t toward ϵ_{min} in dense regions, SAMPPO keeps the Margin-Capacity penalty [Mohri and Zhong, 2026a] from saturating as severely, reducing the gap where PPO's bounds fail (Appendix D).

5 Experiments

We evaluate SAMPPO on two alignment benchmarks to assess preference win rates against PPO and empirical clipping heuristics (BAPO), test cross-dataset hyperparameter transferability, and observe structural n-gram retention (ROUGE-L [Lin, 2004] and BLEU [Papineni et al., 2002]) and entropy stabilization. Beyond a single LLM judge, we report an objective reward-model evaluation (ArmoRM [Wang et al., 2024a]), sequence-level diversity (Distinct-N [Li et al., 2016]), a second independent judge (GPT-4-Turbo), and a direct validation study of the Semantic Isolation Score.

Table 2: Results on 500 held-out prompts across 5 seeds. SAMPPO yields higher ROUGE-L and BLEU scores (scaled 0–100, larger is better), and outperforms both standard PPO and BAPO [Xi et al., 2025] in preference rates. Percentages may not sum to 100% due to rounding.

Dataset	Model	Algorithm	Objective Metrics (0–100) (↑)		Head-to-Head vs. SAMPPO		
			ROUGE-L	BLEU	SAMPPO Win	Baseline Win	Tie
HH-RLHF	Llama-3.1-8B	PPO Baseline	10.84 ± 0.13	1.42 ± 0.04	**57.7%** ± 2.6	39.7% ± 1.9	2.7% ± 1.3
		BAPO Baseline	11.05 ± 0.10	1.49 ± 0.01	**45.0%** ± 2.4	32.7% ± 1.9	22.3% ± 0.8
		SAMPPO	**11.36 ± 0.06**	**1.57 ± 0.05**	–	–	–
	Gemma-2-2B	PPO Baseline	11.83 ± 0.13	1.54 ± 0.03	**55.0%** ± 2.4	37.3% ± 2.3	7.7% ± 0.9
		BAPO Baseline	12.11 ± 0.09	1.60 ± 0.05	**45.7%** ± 2.7	37.8% ± 2.1	16.5% ± 1.1
		SAMPPO	**12.30 ± 0.05**	**1.64 ± 0.03**	–	–	–
HelpSteer	Llama-3.1-8B	PPO Baseline	20.87 ± 0.04	5.54 ± 0.03	**46.6%** ± 2.3	40.3% ± 1.4	13.1% ± 1.1
		BAPO Baseline	20.94 ± 0.08	5.59 ± 0.02	**41.9%** ± 2.4	36.9% ± 1.3	21.2% ± 1.0
		SAMPPO	**21.18 ± 0.02**	**5.67 ± 0.06**	–	–	–
	Gemma-2-2B	PPO Baseline	19.38 ± 0.13	4.62 ± 0.09	**45.9%** ± 1.6	38.2% ± 1.3	15.9% ± 1.2
		BAPO Baseline	19.51 ± 0.11	4.71 ± 0.06	**42.5%** ± 1.7	37.2% ± 1.2	20.3% ± 1.1
		SAMPPO	**19.72 ± 0.09**	**4.81 ± 0.07**	–	–	–

Experimental Setup. To establish that our approach is architecture-agnostic, we evaluate SAMPPO using `Meta-Llama-3.1-8B-Instruct` [Llama Team, AI @ Meta, 2024] and `gemma-2-2b-it` [Gemma Team, Google DeepMind, 2024] on subsets of the Anthropic HH-RLHF [Bai et al., 2022] and NVIDIA HelpSteer [Wang et al., 2024b] benchmarks. We extract 2,000 paired training queries from each dataset and hold out N=500 prompts for the final evaluation; this controlled scale, following Zhou et al. [2023], isolates the critical early-training window where structural instability manifests (standard PPO collapses by step 80, $\approx$ 1.3 epochs; Figure 2) without entangling reward-model distribution shift. For the reward signal, we use the `DeBERTaV3-large` conversational preference model [Köpf et al., 2023]. All evaluations report the mean and standard deviation across 5 random generation seeds; pairwise statistical significance ($p < 0.05$) was assessed via a two-sided paired t-test on per-seed win rates.

Baselines and SAMPPO. We benchmark SAMPPO against standard PPO (ϵ=0.20) and the heuristic BAPO [Xi et al., 2025]; for a fair comparison, both SAMPPO and BAPO use the exact same dynamic clipping bounds ($\epsilon_t \in [0.10, 0.30]$). The Semantic Isolation Score is approximated using the top-k=5 token embeddings with $D_{\text{bound}} = 0.3$. We additionally compare against offline DPO and SimPO, whose hyperparameters replicate the original authors' search grids on a 10% held-out validation set (full tuning protocol, QLoRA hyperparameters, and parity details in Appendix G). All SAMPPO hyperparameters were selected on HH-RLHF/Llama-3.1-8B only and applied *unchanged* to the out-of-distribution HelpSteer and Gemma-2-2B settings; "cross-dataset hyperparameter transferability" refers precisely to this protocol (training always occurs on a subset of the target benchmark).

5.1 Results

All methods are evaluated on N=500 held-out prompts. Head-to-head responses are judged by GPT-4o [OpenAI, 2024] with randomized presentation order. Generative health is assessed via a reward-vs-KL frontier swept over decoding temperatures $T \in \{0.1, 0.4, 0.7, 1.0, 1.2\}$, and we additionally report objective n-gram metrics (BLEU and ROUGE-L) against dataset reference targets.

Table 1: Head-to-head vs. offline DPO and SimPO (Llama-3.1-8B, GPT-4o, N=500).

Dataset	Baseline	SAMPPO Win	Baseline Win	Tie
HH-RLHF	DPO	**53.0%**	45.0%	2.0%
	SimPO	**50.2%**	48.2%	1.6%
HelpSteer	DPO	**47.6%**	44.8%	7.6%
	SimPO	46.4%	**47.2%**	6.4%

Alignment Performance and Head-to-Head Comparison. We first assessed alignment on the primary HH-RLHF dataset, followed by an out-of-distribution evaluation on HelpSteer to test hyperparameter transferability. The values in Table 2 reflect the base SAMPPO configuration using standard input embeddings (an ablation with output embeddings is provided in Section 5.2). As shown in Table 2, SAMPPO demonstrates systematic gains under transferred hyperparameters across both the Llama-3.1-8B and Gemma-2-2B architectures, consistently outperforming both standard PPO and BAPO [Xi et al., 2025] head-to-head. The structural n-gram metrics confirm that this superior preference performance adheres to reference target distributions without linguistic degradation.

Comparison Against Direct Preference Baselines. Table 1 compares SAMPPO against offline DPO and SimPO (Llama-3.1-8B, single typical run). SAMPPO outperforms DPO on both datasets

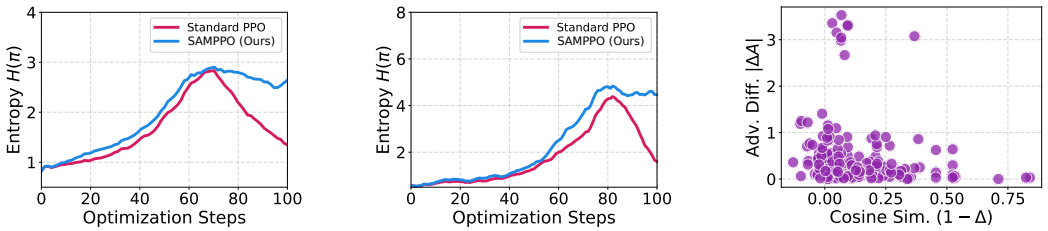


Figure 2: Left & Center: SAMPPO prevents PPO's entropy collapse on HH-RLHF and HelpSteer. Right: Advantage and semantic similarity capture complementary, non-redundant information.

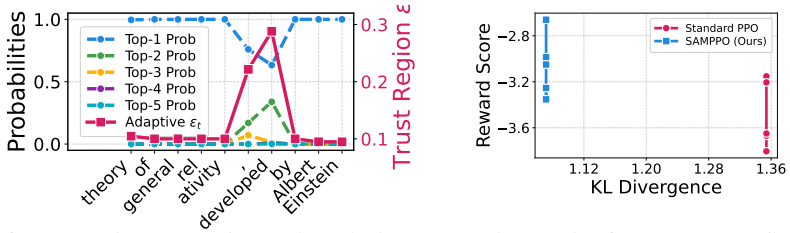


Figure 3: Left: Dynamic trust regions adapt during generation: strict for synonyms, flexible for new concepts. Right: SAMPPO achieves superior rewards at lower KL divergence.

and SimPO on HH-RLHF; the narrow SimPO advantage on HelpSteer is within statistical noise. We do not claim that online RL obsoletes offline methods; rather, once structural instability is fixed, SAMPPO attains parity with tightly tuned, margin-based offline methods while preserving better generative diversity (Figures 2 and 3).

Objective Evaluation Beyond a Single Judge. To rule out judge-specific bias, we re-evaluated the full N=500 held-out generations (Llama-3.1-8B, HH-RLHF) under three additional protocols (Table 4, Appendix J), and SAMPPO is best on every metric: (i) on the *objective* reward model ArmoRM [Wang et al., 2024a], SAMPPO scores 0.0449 vs. 0.0431 (BAPO) and 0.0415 (PPO); (ii) on sequence-level diversity [Li et al., 2016], SAMPPO preserves stable generation diversity (Distinct-2/3: 0.9492/0.9785) versus BAPO (0.8915/0.9320) and PPO, which collapses into repetitive, low-entropy generation (0.8124/0.8540), objectively confirming the mitigation of entropy collapse; and (iii) under a second independent judge, GPT-4-Turbo (temperature 0, 86.4% agreement with GPT-4o), SAMPPO maintains strong wins against both PPO (55.8% vs. 41.2%) and BAPO (53.6% vs. 39.4%).

Entropy Stabilization and Mode Collapse. Tracking the policy entropy $H(\pi)$ during training (Figure 2), standard PPO suffers a sharp entropy collapse around step 80 on HelpSteer (from ~ 4.4 to 1.6) and degrades similarly on HH-RLHF, while SAMPPO mitigates this collapse, plateauing at a stable generative diversity level across both domains, as Theorem 3 predicts.

5.2 Diagnostic Analysis and Ablations

Semantic Isolation Score. A central premise of SAMPPO is that the isolation score captures *contextual semantic density* rather than just policy confidence, token frequency, or embedding anisotropy. A corpus-level study over 28,413 decoding states (500 held-out generations, Llama-3.1-8B, HH-RLHF) confirms: *(i) Not confidence:* Spearman correlation with policy entropy is $\rho = 0.21$ (top-1 probability: -0.17). *(ii) Not frequency:* correlation with pre-training log-frequency is $\rho = -0.12$. *(iii) Synonym ground truth:* WordNet [Miller, 1995] synset labeling yields AUC = 0.94. *(iv) Anisotropy control:* after the Mu and Viswanath [2018] correction, $\rho = 0.93$, AUC = 0.92 (Appendix I).

Orthogonality of Advantage and Distance. The L-shaped distribution of $|\Delta\widehat{A}|$ against semantic similarity $1 - \Delta(a_i, a_j)$ in Figure 2 (right) confirms that Δ and $\widehat{A}$ capture complementary, non-redundant information, justifying their separate roles.

Evolution of ϵ_t. Tracking ϵ_t over a generation trajectory (Figure 3), the trust region tightens toward ~ 0.10 when mass concentrates on a single distinct token and relaxes among distinct conceptual alternatives, allowing exploration.

Hyperparameter and Metric Space Ablations. Table 3 in Appendix H reports sensitivity to k, D_{bound}, and the distance metric (single generation seed, versus the 5-seed averages of Table 2). Performance is robust at k=5; at k=3 the score suffers from high sampling variance and underperforms PPO on HelpSteer, and at k=10 it is diluted by improbable tokens. On HH-RLHF, D_{bound} exhibits an inverted-U confirming 0.3 as optimal; shifting to 0.4 causes underperformance (46.4% vs. 51.8% for PPO), indicating that excessive conservatism in moderately dense regions impedes learning. Using the output unembedding (`lm_head`) achieves the best win rates (60.4% on HH-RLHF); the discrete fallback $\Delta(a, a') = 1 - \delta_{a,a'}$ reduces to baseline levels (47.8%), confirming that the continuous pre-metric is necessary.

Computational Overhead. On an NVIDIA A100, the end-to-end step time (generation, GAE, and 4 PPO epochs) rises from 8.50 s to 8.66 s, a measured $+1.9\%$ overhead with zero extra memory: the top-k cosine computation reuses data already resident in VRAM (Appendix K).

Reward-vs-KL Frontier. Figure 3 shows that, when swept over decoding temperatures, SAMPPO dominates PPO on the Reward-vs-KL frontier at every inference setting, achieving higher rewards at lower KL divergence (comparison against global entropy regularization in Appendix L). We note that this evaluates the quality of the *trained policy* across inference settings, complementing (but not replacing) the standard algorithmic Pareto frontier obtained by sweeping the KL penalty coefficient β across independent training runs.

6 Limitations

We explicitly delineate the scope of our claims. *(i) Token-level proxy:* the pre-metric Δ measures token-embedding distances, a proxy for sequence-level contextual substitutability; our validation study aligns it well with external synonym labels, but the two notions need not coincide in general. *(ii) Theory-implementation gap:* our guarantees concern continuous abstractions (state-level expected divergences, idealized clip conditions), whereas the implementation uses sampled rollouts, approximate advantages, and mini-batch optimization; Propositions 8 and 14 and the budget verification of Section 4.2 bound, but do not eliminate, this dual-approximation gap. In particular, the top-k=5 approximation re-normalizes over a subset of the vocabulary; in high-entropy states where the top-5 tokens hold only a small fraction of the total probability mass, this may overweight the localized contribution and over-tighten the trust region relative to the full-vocabulary expectation assumed in Theorem 7. *(iii) Evaluator dependence:* results depend on the chosen reward model and LLM judges; an objective reward model, diversity metrics, and two independent judges mitigate but cannot exclude residual bias. *(iv) Scale and scope:* experiments use 2,000-query training subsets and N=500 held-out prompts on two conversational benchmarks and two model families; larger scales and additional domains remain future work. *(v) Baseline sensitivity:* despite replicating the original authors' tuning grids under strict parity (Appendix G.1), comparisons retain some sensitivity to hyperparameter selection. *(vi) Failure cases:* for non-semantic tokens (e.g., base64 strings) and domains such as code, math, or multilingual generation, embedding-neighbor geometry may not correspond to semantic (or strict logical) equivalence; the discrete fallback metric ($\Delta(a, a') = 1 - \delta_{a,a'}$) then safely reduces SAMPPO to a probability-weighted trust-region constraint. *(vii) Distributional vs. logical similarity:* token embeddings capture distributional similarity, meaning logically opposing tokens (e.g., "increase" vs. "decrease") may be close in embedding space; in such cases SAMPPO conservatively tightens the trust region, which may slow learning when the advantage signal is already strong and valid.

7 Conclusion

We showed that uniform trust regions are prone to entropy collapse in semantically dense action spaces, even under exchangeable noise, and introduced SAMPPO to resolve this via adaptive, density-governed clipping, grounded in structure-aware monotonic improvement, a provable uniform weighted-KL budget, mirror descent convergence, and $\mathcal{H}$-consistency theory. Empirically, SAMPPO outperforms PPO and BAPO across two independent judges, an objective reward model, and diversity metrics, transferring its hyperparameters unchanged across benchmarks and architectures at a $+1.9\%$ overhead while remaining highly competitive with DPO and SimPO. Future directions include multi-modal action spaces and offline margin-based variants of the structure-aware divergence.

References

J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31, 2017.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. Multi-class $\mathcal{H}$-consistency bounds. In *Advances in Neural Information Processing Systems*, 2022a.

P. Awasthi, A. Mao, M. Mohri, and Y. Zhong. $\mathcal{H}$-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pages 1117–1174, 2022b.

M. G. Azar, Z. D. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455, 2024.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

G. Chen, Y. Peng, and M. Zhang. An adaptive clipping approach for proximal policy optimization. *arXiv preprint arXiv:1804.06461*, 2018.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

C. Cortes, A. Mao, C. Mohri, M. Mohri, and Y. Zhong. Cardinality-aware set prediction and top-k classification. In *Advances in Neural Information Processing Systems*, 2024.

C. Cortes, A. Mao, M. Mohri, and Y. Zhong. Balancing the scales: A theoretical and algorithmic framework for learning from imbalanced data. In *International Conference on Machine Learning*, 2025a.

C. Cortes, M. Mohri, and Y. Zhong. Improved balanced classification with theoretically grounded loss functions. In *Advances in Neural Information Processing Systems*, 2025b.

C. Cortes, A. Mao, M. Mohri, and Y. Zhong. Optimized deferral for imbalanced settings. In *International Conference on Machine Learning*, 2026a.

C. Cortes, M. Mohri, and Y. Zhong. A theoretical framework for modular learning of robust generative models. In *International Conference on Machine Learning*, 2026b.

G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen, Z. Liu, H. Peng, L. Bai, W. Ouyang, Y. Cheng, B. Zhou, and N. Ding. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

G. DeSalvo, C. Mohri, M. Mohri, and Y. Zhong. Budgeted multiple-expert deferral. *arXiv preprint arXiv:2510.26706*, 2025.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115, 2023.

K. Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65, 2019.

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. KTO: Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, 2024.

C. Gao, C. Zheng, X.-H. Chen, K. Dang, S. Liu, B. Yu, A. Yang, S. Bai, J. Zhou, and J. Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

P. He, J. Gao, and W. Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, N. M. Duc, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. Mattick. OpenAssistant conversations – democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681, 2023.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, 2016.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

A. Mao. *Theory and Algorithms for Learning with Multi-Class Abstention and Multi-Expert Deferral*. PhD thesis, New York University, 2025.

A. Mao, C. Mohri, M. Mohri, and Y. Zhong. Two-stage learning to defer with multiple experts. In *Advances in Neural Information Processing Systems*, 2023a.

A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds: Characterization and extensions. In *Advances in Neural Information Processing Systems*, 2023b.

A. Mao, M. Mohri, and Y. Zhong. Ranking with abstention. In *ICML 2023 Workshop on The Many Facets of Preference-Based Learning*, 2023c.

A. Mao, M. Mohri, and Y. Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, 2023d.

A. Mao, M. Mohri, and Y. Zhong. H-consistency bounds for pairwise misranking loss surrogates. In *International Conference on Machine Learning*, pages 23743–23802, 2023e.

A. Mao, M. Mohri, and Y. Zhong. Structured prediction with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, volume 36, pages 46903–46937, 2023f.

A. Mao, M. Mohri, and Y. Zhong. Principled approaches for learning to defer with multiple experts. In *International Symposium on Artificial Intelligence and Mathematics*, 2024a.

A. Mao, M. Mohri, and Y. Zhong. Predictor-rejector multi-class abstention: Theoretical analysis and algorithms. In *International Conference on Algorithmic Learning Theory*, 2024b.

A. Mao, M. Mohri, and Y. Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, 2024c.

A. Mao, M. Mohri, and Y. Zhong. H-consistency guarantees for regression. In *International Conference on Machine Learning*, pages 34712–34737, 2024d.

A. Mao, M. Mohri, and Y. Zhong. Multi-label learning with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, 2024e.

A. Mao, M. Mohri, and Y. Zhong. Realizable H-consistent and Bayes-consistent loss functions for learning to defer. In *Advances in Neural Information Processing Systems*, 2024f.

A. Mao, M. Mohri, and Y. Zhong. Regression with multi-expert deferral. In *International Conference on Machine Learning*, pages 34738–34759, 2024g.

A. Mao, M. Mohri, and Y. Zhong. A universal growth rate for learning with smooth surrogate losses. In *Advances in Neural Information Processing Systems*, 2024h.

A. Mao, M. Mohri, and Y. Zhong. Mastering multiple-expert routing: Realizable H-consistency and strong guarantees for learning to defer. In *International Conference on Machine Learning*, 2025a.

A. Mao, M. Mohri, and Y. Zhong. Principled algorithms for optimizing generalized metrics in binary classification. In *International Conference on Machine Learning*, 2025b.

A. Mao, M. Mohri, and Y. Zhong. Enhanced H-consistency bounds. In *International Conference on Algorithmic Learning Theory*, 2025c.

Y. Meng, M. Xia, and D. Chen. SimPO: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

C. Mohri, D. Andor, E. Choi, M. Collins, A. Mao, and Y. Zhong. Learning to reject with a fixed predictor: Application to decontextualization. In *International Conference on Learning Representations*, 2024.

M. Mohri and Y. Zhong. Mind the gap: Structure-aware consistency in preference learning. In *International Conference on Machine Learning*, 2026a.

M. Mohri and Y. Zhong. Linear-core surrogates: Smooth loss functions with linear rates for classification and structured prediction. In *International Conference on Machine Learning*, 2026b.

M. Mohri and Y. Zhong. Beyond Tsybakov: Model margin noise and H-consistency bounds. In *International Symposium on Artificial Intelligence and Mathematics*, 2026c.

M. Mohri and Y. Zhong. Principled algorithms for optimizing generalized metrics in multi-label learning. *arXiv preprint arXiv:2605.28767*, 2026d.

M. Mohri, J. Schneider, and Y. Zhong. Generalized distributional alignment games for unbiased answer-level fine-tuning. *arXiv preprint arXiv:2605.02435*, 2026.

J. Mu and P. Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

X. Qiu and R. Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in Neural Information Processing Systems*, 37:134507–134533, 2024.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

B. Rahman. PPO-BR: dual-signal entropy-reward adaptation for trust region policy optimization. *arXiv preprint arXiv:2505.17714*, 2025.

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec. TRL: Transformer Reinforcement Learning. https://github.com/huggingface/trl, 2020.

H. Wang, W. Xiong, T. Xie, H. Zhao, and T. Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024a.

Y. Wang, H. He, X. Tan, and Y. Gan. Trust region-guided proximal policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023.

Z. Wang, Y. Dong, J. Zeng, V. Adams, M. N. Sreedhar, D. Egert, O. Delalleau, J. Scowcroft, N. Kant, A. Swope, and O. Kuchaiev. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3371–3384, 2024b.

Z. Xi, X. Guo, Y. Nan, E. Zhou, J. Shen, W. Chen, J. Liu, J. Huang, Z. Zhang, H. Guo, X. Deng, Z. Lei, M. Zheng, G. Wang, S. Zhang, P. Sun, R. Zheng, H. Yan, T. Gui, Q. Zhang, and X. Huang. BAPO: stabilizing off-policy reinforcement learning for LLMs via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*, 2025.

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, X. Liu, H. Lin, Z. Lin, B. Ma, G. Sheng, Y. Tong, C. Zhang, M. Zhang, W. Zhang, H. Zhu, J. Zhu, J. Chen, J. Chen, C. Wang, H. Yu, Y. Song, X. Wei, H. Zhou, J. Liu, W.-Y. Ma, Y.-Q. Zhang, L. Yan, M. Qiao, Y. Wu, and M. Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Y. Zhong. *Fundamental Novel Consistency Theory: H-Consistency Bounds*. PhD thesis, New York University, 2025.

C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2023.

Contents of Appendix

A Related Work **16**

B Notes on Action Space Geometry and Semantic Isolation **17**

C Algorithmic Comments **17**

C.1 Core Algorithmic Modifications 17

C.2 Revisiting the Entropy Collapse Example 18

D Theoretical Foundations: Structure-Aware Improvement **18**

D.1 Structure-Aware Divergence 18

D.2 Structure-Aware Performance Bound 19

D.3 Adaptive Trust Region Interpretation 19

D.4 Mirror Descent Interpretation 19

D.5 Convergence Analysis via State-Dependent Mirror Descent 20

D.6 Connection to Margin-Shifted $\mathcal{H}$-Consistency 21

D.7 Discussion 22

E Omitted Proofs **22**

E.1 Proof of Theorem 3 22

E.2 Proof of Corollary 4 23

E.3 Proof of Lemma 6 24

E.4 Proof of Theorem 7 24

E.5 Proof of Proposition 8 and Supporting Remarks 25

E.5.1 Empirical Verification of the KL Budget 26

E.6 Proof of Lemma 11 26

E.7 Proof of Theorem 9 27

E.8 Proof of Theorem 10 28

F Supporting Theorems for $\mathcal{H}$-Consistency **30**

G Implementation Details of SAMPPO and Baselines **31**

G.1 Baseline Tuning Protocol 32

H Ablation Study **33**

I Validation Study of the Semantic Isolation Score **33**

I.1 Qualitative Examples of Decoding States 33

I.2 Corpus-Level Systematic Analysis 34

J Extended Objective Evaluation **34**

K Wall-Clock Profiling **35**

L SAMPPO vs. Global Entropy Regularization 35

M Broader Societal Impacts 35

A Related Work

Our work intersects several key areas of preference learning and optimization. For comprehensive overviews of LLM alignment and its fundamental limitations, we refer readers to recent surveys [Wang et al., 2023, Casper et al., 2023].

Reinforcement Learning from Human Feedback (RLHF): The standard RLHF paradigm fine-tunes language models by adopting PPO against a learned reward model [Christiano et al., 2017, Stiennon et al., 2020, Ouyang et al., 2022, Bai et al., 2022]. While highly effective and deployed in scalable open-source systems such as DAPO [Yu et al., 2025], PPO is known to be sensitive to hyperparameters and reward noise, making it susceptible to over-optimization and generative degradation. Recent studies are increasingly investigating the underlying mechanics of these instabilities, such as the dynamics of the entropy mechanism in reasoning language models [Cui et al., 2025].

Direct Preference Alignment: Following the introduction of DPO [Rafailov et al., 2023], which bypasses explicit reward modeling, several offline variations have emerged to address theoretical and empirical limitations. IPO [Azar et al., 2024] incorporates a squared loss to prevent overfitting. KTO [Ethayarajh et al., 2024] applies prospect theory to align models without paired preference data. Furthermore, SimPO [Meng et al., 2024] and SLiC-HF [Zhao et al., 2023] demonstrate the empirical benefits of incorporating hard margins into their respective objectives. While scalable, these methods optimize fixed divergence penalties. Our work shows that online RL, when equipped with structure-aware mechanisms, remains highly competitive with these tuned offline methods.

Adaptive Trust Regions and Entropy Collapse: Standard PPO uses a uniform clipping hyperparameter to enforce the trust region. Adaptive clipping mechanisms based on learning progress [Chen et al., 2018] or theoretical KL targets [Wang et al., 2019] have been explored previously in continuous control. In the constrained RL setting, Achiam et al. [2017] derive performance-difference bounds in which the divergence penalty is a total variation distance *averaged* over the state distribution rather than a worst-case divergence; Theorem 7(c) leverages this result to make our average-case structure-aware surrogate rigorous. More recently in LLM alignment, Soft Adaptive Policy Optimization (SAPO) [Gao et al., 2025] dynamically adjusts constraints via continuous advantage-based gates, and PPO-BR [Rahman, 2025] introduces dual-signal entropy-reward adaptation. Most closely related to our work is BAPO [Xi et al., 2025], which similarly identifies the phenomenon of "entropy collapse" during LLM fine-tuning and proposes an adaptive clipping mechanism. However, their algorithm determines clipping bounds via an empirical search heuristic to balance the policy and lacks theoretical or convergence analysis. In contrast, SAMPPO modulates the trust region based on the geometric semantic topology of the action space. Thus, our structure-aware monotonic improvement theory and consistency bounds can provide a formal mathematical foundation for the empirical observations in BAPO.

Consistency Theory and Semantic Density: The consistency of surrogate losses is a foundational problem in statistical learning [Zhang, 2004, Bartlett et al., 2006]. Recent advancements in $\mathcal{H}$-consistency bounds specifically address restricted hypothesis sets like neural networks [Awasthi et al., 2022a,b, Mao et al., 2023b, 2025c, Mohri and Zhong, 2026c], with a growing body of follow-up work: general tools such as smooth and cross-entropy surrogates, growth rates, and linear-rate surrogates [Mao et al., 2023d, 2024h, Mohri and Zhong, 2026b]; extensions to structured prediction, ranking, regression, multi-label learning, and generalized metrics [Mao et al., 2023f,e,c, 2024d,e, Cortes et al., 2024, Mao et al., 2025b, Mohri and Zhong, 2026d]; class-imbalanced learning [Cortes et al., 2025a,b]; abstention and learning to defer [Mao et al., 2024c,b, Mohri et al., 2024, Mao et al., 2023a, 2024a,f,g, 2025a, Cortes et al., 2026a, DeSalvo et al., 2025]; and generative modeling and alignment [Cortes et al., 2026b, Mohri et al., 2026]; see also the theses of Mao [2025] and Zhong [2025]. In preference learning, Mohri and Zhong [2026a] prove that enforcing a separation margin is a strict requirement for consistency, introducing Structure-Aware DPO (SA-DPO) which scales offline margins based on semantic distance. Inspired by this framework, we extend structure-aware consistency into the online reinforcement learning regime. Furthermore, our geometrically-grounded evaluation connects directly to recent literature using semantic density for uncertainty quantification

[Qiu and Miikkulainen, 2024], which demonstrates that structural geometry in semantic space reliably captures model confidence and ambiguity.

B Notes on Action Space Geometry and Semantic Isolation

As empirically demonstrated in Section 5.2, rather than using static input embeddings, evaluating the cosine distance between the output unembedding (`lm_head`) vectors better captures the model's actual semantic separation capacity.

The Semantic Distance and the estimated Advantage $\widehat{A}(s,a)$ represent different aspects of the learning problem. Advantage represents *directional value* (how aligned the token is with human intent), whereas Semantic Distance represents *structural geometry* (how distinct the token is from its neighbors). Both play an important role in our proposed algorithm.

Note that if the action space lacks a continuous semantic topology (e.g., discrete categorical choices or missing embedding architectures), Δ defaults gracefully to the discrete metric: $\Delta(a,a') = 1-\delta_{a,a'}$, where $\delta_{a,a'}$ is the Kronecker delta (yielding 1 if $a = a'$, and 0 otherwise). Under this fallback metric, the expected semantic isolation reduces to $D_{\pi_{\text{old}}}(s,a) = 1 - \pi_{\text{old}}(a|s)$. This automatically reduces SAMPPO to a probability-weighted trust-region constraint, dynamically shrinking the trust region for actions with high policy certainty.

C Algorithmic Comments

C.1 Core Algorithmic Modifications

SAMPPO enhances the standard PPO framework through three primary components designed to measure and respond to local action density. To ensure mathematical stationarity and eliminate inner-loop computational needs, these components are evaluated directly from the fixed behavior policy π_{old} prior to the optimization epoch loop.

State-Dependent Theory vs. Action-Dependent Implementation. While our subsequent theoretical analysis (Section 4) constructs a state-dependent expected divergence abstraction to prove global convergence, our practical SAMPPO algorithm applies an *action-dependent* clipping bound ϵ_t. By computing the bound using the specific token a_t sampled at timestep t, SAMPPO provides a precise sample-level trust region while bypassing the computational overhead of evaluating the expectation across the full vocabulary.

Action Distance Pre-Metric, $\Delta(a,a')$: We first require a computable, bounded pre-metric to quantify the dissimilarity between any two actions. For natural language tasks, we typically use the cosine distance between action embeddings, $\Delta(a,a') = 1 - \cos(E(a), E(a'))$.

Semantic Isolation Score, $D_{\pi_{old}}(s_t,a_t)$: To determine the appropriate trust region size, we calculate a Semantic Isolation Score (Eq. 1), which measures the distinctiveness of the chosen action a_t relative to other probable actions. To ensure a stationary surrogate objective and prevent gradient routing through the clipping bounds, this expectation is evaluated with respect to the *fixed behavior policy* π_{old}.

In practice, to avoid the stochastic variance associated with Monte Carlo sampling while bypassing the computational overhead of evaluating the full vocabulary, this expectation is approximated deterministically using the top-k most probable actions $\{a'_i\}_{i=1}^{k}$ calculated *once* per rollout step. This effectively evaluates the conditional expectation restricted to the plausible linguistic head.

Adaptive Clipping Mechanism, ϵ_t: Rather than a fixed global value, SAMPPO computes a per-timestep clipping range ϵ_t. This value is modulated between the hyperparameters $\epsilon_{\min}$ and $\epsilon_{\max}$ based on the normalized isolation score, incorporated in the scaling factor S_t:

$$\epsilon_t = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot S_t. \tag{14}$$

To avoid the instability of batch-level normalization across small mini-batches, we use an absolute metric bound D_{bound}, setting $S_t = \min(D_{\pi_{\text{old}}}(s_t,a_t)/D_{\text{bound}}, 1.0)$. This acts as a bounded normalization, mapping scores to $[0,1]$, and because ϵ_t is computed from the fixed behavior policy prior to the epoch loop, it acts as a detached scalar constant (via a `stop_gradient` operation). Finally, because

the scaling factor is detached from the computation graph, the distance function $\Delta(a, a')$ does not need to be differentiable with respect to the policy parameters.

C.2 Revisiting the Entropy Collapse Example

To demonstrate how SAMPPO lessens policy collapse, we revisit the synonym cluster example from Section 2.2. Recall that for a cluster of κ indistinguishable synonyms, standard PPO suffers an entropy drop proportional to $\frac{\kappa-1}{2}\epsilon^2$ in the small-cluster regime.

In SAMPPO, because the κ synonyms are semantically identical, the distance between them approaches zero ($\Delta(a_i, a_j) \approx 0$). Therefore, the Semantic Isolation Score for any token in the cluster collapses ($D_{\pi_{\text{old}}}(s, a_i) \to 0$). The adaptive clipping mechanism (Eq. 14) then tightly bounds the trust region, driving $\epsilon_t \to \epsilon_{\min}$.

Substituting this dynamic bound back into Theorem 3, the resulting entropy drop is approximately bounded by $\frac{\kappa-1}{2}\epsilon_{\min}^2$. By selectively restricting the trust region in dense semantic neighborhoods, SAMPPO achieves a quadratic reduction in entropy collapse, while still retaining the ability to take large exploratory steps ($\epsilon_{\max}$) in semantically sparse regions.

D Theoretical Foundations: Structure-Aware Improvement

This section establishes a principled theoretical foundation for SAMPPO. We show that by introducing a structure-aware, state-weighted divergence, one can derive a valid performance improvement bound analogous to those used in trust-region methods. This formulation preserves theoretical correctness while capturing the non-uniform geometry of the action space.

D.1 Structure-Aware Divergence

In standard trust-region methods, policy updates are controlled via a global divergence such as the Kullback–Leibler (KL) divergence. However, this uniform measure does not account for the fact that, in structured action spaces such as natural language, different regions exhibit varying semantic density. To address this, we define a structure-aware expected weighted divergence.

Definition 5 (Structure-Aware Weighted KL Divergence)**.** *Let $D_{\pi_{old}}(s, a)$ denote the Semantic Isolation Score and define the expected state-level isolation: $D(s) = \mathbb{E}_{a\sim\pi_{old}(\cdot|s)}[D_{\pi_{old}}(s, a)]$. For a fixed stability constant $\nu > 0$, define the weight: $w(s) = \frac{1}{D(s)+\nu}$. Let $\rho_{\pi_{old}}$ denote the state visitation distribution induced by π_{old}. The structure-aware divergence between policies is defined via the KL divergence as:*

$$D_w(\pi, \pi_{old}) = \mathop{\mathbb{E}}_{s\sim\rho_{\pi_{old}}}\Big[w(s)\mathsf{D}_{\mathrm{KL}}\big(\pi(\cdot|s) \parallel \pi_{old}(\cdot|s)\big)\Big]. \tag{5}$$

The additive constant ν ensures that the weights remain bounded, preventing the divergence from exploding in regions of extremely high semantic density. The following lemma establishes that this divergence is equivalent, up to bounded constants, to the standard average KL divergence, ensuring it is a well-behaved metric for optimization.

Lemma 6 (Equivalence to Average KL)**.** *Let the state-level isolation score $D(s)$ be bounded such that $0 \le D_{\min} \le D(s) \le D_{\max}$. Let the weight function be $w(s) = \frac{1}{D(s)+\nu}$ for some constant $\nu > 0$. The structure-aware divergence $D_w(\pi, \pi_{old})$ and the standard average KL divergence $\mathsf{D}_{\mathrm{KL}}^{avg}$ satisfy the following two-sided inequality:*

$$(D_{\min} + \nu) \cdot D_w(\pi, \pi_{old}) \le \mathsf{D}_{\mathrm{KL}}^{avg} \le (D_{\max} + \nu) \cdot D_w(\pi, \pi_{old}). \tag{6}$$

Proof Sketch. The proof proceeds by substituting the bounds $D_{\min}$ and $D_{\max}$ into the relationship between the structure-aware weight $w(s)$ and the standard average KL divergence. This gives a direct pointwise bound on the expectation, establishing the two-sided inequality. The full proof is given in Appendix E.3. □

D.2 Structure-Aware Performance Bound

We now establish a monotonic improvement bound using the structure-aware divergence. This justifies why the adaptive trust region in SAMPPO preserves performance while preventing synonym collapse.

Theorem 7 (Structure-Aware Surrogate Improvement)**.** *Let* $C_0 = \frac{2\gamma}{(1-\gamma)^2} \max_{s,a} |A_{\pi_{old}}(s,a)|$ *and* $C = C_0 \cdot (D_{\max} + \nu)$*, where* $J(\pi)$ *denotes the expected return.*

(a) ***Rigorous bound.*** *The structure-aware surrogate objective satisfies:*

$$J(\pi) \geq J(\pi_{old}) + L_{\pi_{old}}(\pi) - C \cdot \max_s \Big[w(s) \mathsf{D}_{\mathrm{KL}}\big(\pi(\cdot|s) \parallel \pi_{old}(\cdot|s)\big)\Big]. \tag{7}$$

(b) ***Practical surrogate.*** *As in Schulman et al. [2015], replacing the intractable worst-case maximum with the expected value under* $\rho_{\pi_{old}}$ *yields the surrogate used in practice:*

$$\widetilde{J}(\pi) = J(\pi_{old}) + L_{\pi_{old}}(\pi) - C \cdot D_w(\pi, \pi_{old}). \tag{8}$$

Since $\mathbb{E}_s[\cdot] \leq \max_s[\cdot]$*, the surrogate* $\widetilde{J}(\pi)$ *upper-bounds the rigorous bound (7); it is therefore a less conservative (but still consistent) proxy for policy improvement.*

(c) ***Rigorous average-divergence bound.*** *Let* $\zeta^\pi = \max_s \big|\mathbb{E}_{a\sim\pi(\cdot|s)}[A_{\pi_{old}}(s,a)]\big|$*, so that* $\frac{2\gamma\zeta^\pi}{(1-\gamma)^2} \leq C_0$*. Then*

$$J(\pi) \geq J(\pi_{old}) + L_{\pi_{old}}(\pi) - \frac{2\gamma\zeta^\pi}{(1-\gamma)^2} \sqrt{\frac{D_{\max} + \nu}{2} D_w(\pi, \pi_{old})}. \tag{9}$$

Proof Sketch. Using the standard trust-region bound, the performance difference is bounded by the maximum KL divergence across all states. We bound the unweighted pointwise KL divergence by $(D_{\max} + \nu)\, w(s)\, \mathsf{D}_{\mathrm{KL}}$, which scales the penalty by the structure-aware divergence, yielding part (a). For part (b), following Schulman et al. [2015], we replace the intractable worst-case maximum with the tractable expectation D_w, obtaining the practical surrogate. Part (c) makes this average-case penalty rigorous via the state-averaged total-variation performance bound of Achiam et al. [2017]. The full proof is deferred to Appendix E.4. □

D.3 Adaptive Trust Region Interpretation

We now interpret SAMPPO as implementing a state-action-dependent trust-region heuristic. Recall that SAMPPO uses a per-timestep clipping parameter $\epsilon_t = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot S_t$, where S_t is a normalized scaling function computed deterministically via the reference policy. This mechanism induces the following localized behavior:

Semantically dense regions (low $D_{\pi_{old}}(s,a)$*):* $\epsilon_t \to \epsilon_{\min}$. Updates are tightly constrained, preventing the exploitation of noisy signals among indistinguishable actions.

Semantically sparse regions (high $D_{\pi_{old}}(s,a)$*):* $\epsilon_t \to \epsilon_{\max}$. Larger updates are permitted, allowing the model to learn efficiently when a distinctly superior action is identified.

Thus, SAMPPO adaptively allocates the "divergence budget" across states, placing stricter constraints in regions where policy changes are most likely to overfit noise. While this mechanism does not explicitly enforce a hard constraint on D_w, it provides a practical approximation to a structure-aware trust region by modulating update magnitudes in accordance with semantic density.

D.4 Mirror Descent Interpretation

Conceptually, standard policy optimization can be viewed as mirror descent in policy space. The Bregman divergence generated by negative entropy $\Psi(\pi) = \sum_a \pi(a|s) \log \pi(a|s)$ uniquely yields the KL divergence $\mathsf{D}_{\mathrm{KL}}(\pi \| \pi_{\mathrm{old}})$, with the new policy as the first argument, rather than $\mathsf{D}_{\mathrm{KL}}(\pi_{\mathrm{old}} \| \pi)$ as in the original TRPO formulation [Schulman et al., 2015]. The Surrogate Improvement bound (Theorem 7) remains valid because Pinsker's inequality $D_{\mathrm{TV}}^2(p,q) \leq \frac{1}{2}\mathsf{D}_{\mathrm{KL}}(p\|q)$ holds for either argument order and the total variation distance is symmetric. In contrast to PPO, SAMPPO replaces the uniform KL divergence with our structure-aware weighted divergence D_w. This induces a *state-dependent Bregman divergence* on the policy space, leading to the modified mirror descent update for

the policy π_k:

$$\pi_{k+1} = \underset{\pi}{\operatorname{argmax}} \left\{ \underset{s,a\sim\pi_k}{\mathbb{E}} \left[A_k(s,a) \frac{\pi(a|s)}{\pi_k(a|s)} \right] - \frac{1}{\eta} D_{w_k}(\pi, \pi_k) \right\}. \tag{15}$$

This perspective provides a principled connection to geometry-aware optimization. Note that under standard first-order policy gradient approximations, the inner product $\langle \nabla J(\pi_k), \pi \rangle$ in the mirror descent update is equivalent to the expected surrogate advantage $\mathbb{E}_{s,a\sim\pi_k}[A_k(s,a)\pi(a|s)/\pi_k(a|s)]$, bridging the abstract formulation with the implementable surrogate. Under a local quadratic approximation around the policy parameters θ_k, mirror descent with a KL divergence is equivalent to natural policy gradient (NPG) ascent. The standard NPG update for the parameters θ of a policy π_θ is given by:

$$\theta_{k+1} \approx \theta_k + \eta F(\theta_k)^{-1} \nabla_\theta L(\theta_k), \tag{16}$$

where $L(\theta_k)$ is the policy gradient objective and $F(\theta_k)$ is the Fisher Information Matrix (FIM), which defines the natural geometry of the policy manifold.

The weighted divergence D_{w_k} used by SAMPPO corresponds to modifying this geometry. Specifically, it induces a state-weighted Fisher Information Matrix:

$$F_{w_k}(\theta_k) = \underset{s\sim\rho_{\pi_k}}{\mathbb{E}} \left[w_k(s) F_s(\theta_k) \right], \tag{17}$$

where $F_s(\theta_k)$ is the per-state FIM. The optimization of SAMPPO's objective is therefore implicitly taking steps that approximate a semantically-aware natural gradient ascent:

$$\theta_{k+1} \approx \theta_k + \eta F_{w_k}(\theta_k)^{-1} \nabla_\theta L(\theta_k). \tag{18}$$

In semantically dense regions, $w_k(s)$ is large, which up-weights the FIM, increases the local metric, and results in smaller, more cautious steps. In sparse regions, $w_k(s)$ is small, permitting larger steps. This interpretation shows that SAMPPO is not merely a heuristic clipping strategy, but can be understood as a form of *adaptive preconditioning* where the gradient is scaled according to the semantic structure of the action space.

D.5 Convergence Analysis via State-Dependent Mirror Descent

To transcend the "one-step" monotonic improvement of Theorem 7 and provide a formal global convergence guarantee, we first establish that this structure-aware divergence provides a valid geometry for mirror descent.

Lemma 11 (Structure-Aware Bregman Divergence)**.** *Let the step-dependent structure-aware divergence for mirror descent be defined via the KL divergence:* $D_{w_k}(\pi, \pi_k) = \mathbb{E}_{s\sim\rho_{\pi_k}}[w_k(s)\mathrm{D}_{\mathrm{KL}}(\pi(\cdot|s) \parallel \pi_k(\cdot|s))]$*. This is a valid Bregman divergence generated by the state-weighted negative entropy:*

$$\Psi_{w_k}(\pi) = \underset{s\sim\rho_{\pi_k}}{\mathbb{E}} \left[w_k(s) \sum_a \pi(a|s) \log \pi(a|s) \right]. \tag{19}$$

Furthermore, D_{w_k} *is strongly convex with respect to the expected squared* ℓ_1*-norm (*$\mathbb{E}_s[\|\pi(\cdot|s) - \pi_k(\cdot|s)\|_1^2]$*) with a uniform global modulus* $\mu = \frac{1}{D_{\max}+\nu} > 0$*.*

Proof Sketch. Since the weight $w_k(s)$ does not depend on the optimization variable π, the state-wise Bregman divergences generated by the negative entropy match the weighted KL divergence exactly. Uniform strong convexity with modulus μ is derived using Pinsker's inequality and the bound $w_k(s) \geq \frac{1}{D_{\max}+\nu}$. The complete proof is in Appendix E.6. □

Since D_{w_k} provides a valid strongly convex Bregman divergence that maintains a uniform global lower bound μ across iterations, the SAMPPO update provides a formal global convergence guarantee.

Theorem 9 (Convergence to a Stationary Point)**.** *Assume the expected policy objective* $J(\pi)$ *is* L*-smooth uniformly across shifting state distributions* ρ_{π_k} *with respect to the expected squared* ℓ_1*-norm (*$\mathbb{E}_s[\|\pi(\cdot|s) - \pi_k(\cdot|s)\|_1^2]$*), implicitly absorbing a worst-case concentrability coefficient standard in*

MDP analysis, and bounded above by J^*. Let the sequence of policies $\{\pi_k\}_{k=0}^T$ be generated by the SAMPPO mirror descent update:

$$\pi_{k+1} = \operatorname*{argmax}_{\pi} \left\{ \langle \nabla J(\pi_k), \pi \rangle - \frac{1}{\eta} D_{w_k}(\pi, \pi_k) \right\}. \tag{10}$$

For a step size $\eta \leq \frac{\mu}{L}$, SAMPPO converges to a stationary point at a rate of $\mathcal{O}(1/T)$:

$$\min_{0 \leq k < T} \mathop{\mathbb{E}}_{s} \left[\|\pi_{k+1} - \pi_k\|_1^2 \right] \leq \frac{2\eta(D_{\max} + \nu)}{T} (J^* - J(\pi_0)). \tag{11}$$

Proof Sketch. By evaluating the first-order optimality condition of the mirror descent update, we bound the policy improvement using the symmetrized Bregman divergence. The uniform μ-strong convexity of D_{w_k} ensures a strictly positive descent step coefficient $\frac{\mu}{2\eta}$. Telescoping the bounded L-smooth objective over T iterations yields the global $\mathcal{O}(1/T)$ convergence rate. The full proof is detailed in Appendix E.7. □

This result confirms that SAMPPO is not merely a stabilizing heuristic, but a principled optimization algorithm. In semantically dense regions, the local semantic weight $w_k(s)$ becomes large, which effectively scales up the local strong convexity of the divergence penalty. This acts as an adaptive preconditioner that forces smaller, more conservative steps in ambiguous regions, ensuring stable global convergence.

D.6 Connection to Margin-Shifted $\mathcal{H}$-Consistency

We further ground SAMPPO's adaptive clipping mechanism in statistical learning theory. Recent advancements in $\mathcal{H}$-consistency for preference learning demonstrate that for restricted equicontinuous hypothesis sets (e.g., neural networks), standard unconstrained surrogate losses yield vacuous generalization bounds. To achieve theoretical consistency, algorithms must enforce a separation margin γ, optimizing a *margin-shifted surrogate* $\Phi_\gamma(u) = \Phi(u - \gamma)$. Definitions for the specific bounds and approximation gaps are formally provided in Appendix F.

In the offline direct preference setting, prior work resolves this by proposing Structure-Aware DPO (SA-DPO) [Mohri and Zhong, 2026a], which applies a dynamic pair-dependent margin $\Gamma(y, y') = \tau\Delta(y, y')$. We now show that SAMPPO's adaptive trust region acts as the online reinforcement learning counterpart to this mechanism. By bounding the surrogate update, SAMPPO implicitly optimizes a structure-aware margin-shifted surrogate, satisfying the requirements for Structure-Aware $\mathcal{H}$-Consistency.

Theorem 10 (Implicit Structure-Aware $\mathcal{H}$-Consistency)**.** *Let $h_\theta(s, a) = \log \pi_\theta(a|s) - \log \pi_{\text{old}}(a|s)$ be the relative scoring function of the policy hypothesis set $\mathcal{H}$. Up to a second-order approximation error uniformly bounded within the trust region (as explicitly quantified in Proposition 14), the SAMPPO clipped objective is mathematically equivalent to minimizing an advantage-weighted structure-aware margin-shifted Hinge surrogate:*

$$\mathsf{L}_{\Phi,\Gamma}(h_\theta) = |A_t| \Phi_{\text{hinge}}(y_t \cdot h_\theta(s_t, a_t) - \Gamma(s_t, a_t)), \tag{12}$$

where $y_t = \operatorname{sign}(A_t)$, $\Phi_{\text{hinge}}(u) = \max(0, -u)$, and the local margin is $\Gamma(s_t, a_t) = \epsilon_t$. Thus, SAMPPO satisfies the established structure-aware $\mathcal{H}$-consistency bound:

$$\mathcal{R}(h_\theta) - \mathcal{R}^*(\mathcal{H}) + \mathcal{M}(\mathcal{H}) \leq \mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(h_\theta) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H}) + \mathcal{A}_\Gamma(\mathcal{H}), \tag{13}$$

where $\mathcal{R}(h_\theta)$ is the target 0-1 ranking error, $\mathcal{M}(\mathcal{H})$ is the minimizability gap, $\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma} = \mathsf{L}_{\Phi_{\text{hinge}},\Gamma}/\Phi_{\text{hinge}}(-\Gamma)$ is the inverse-margin weighted surrogate loss, and $\mathcal{A}_\Gamma(\mathcal{H}) = \mathbb{E}[\mathcal{C}^*_{\mathsf{L}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H})] - \mathbb{E}[\mathcal{C}^*(\mathcal{H})]$ is the structure-aware approximation gap.

Proof Sketch. A local first-order Taylor expansion of the probability ratio around the clipping bounds demonstrates that, for both positive and negative advantages, the optimization reduces to an advantage-weighted, margin-shifted Hinge surrogate. Since the local margin $\Gamma(s_t, a_t) = \epsilon_t$ scales proportionally to the local semantic isolation and its gradient is statically detached, it fulfills all structural requirements for $\mathcal{H}$-consistency. The detailed proof is in Appendix E.8, where Proposition 14 further quantifies the approximation error. □

We now analyze why SAMPPO prevents policy collapse by mapping its dynamic constraints to the *Margin-Capacity Profile*. As detailed in Appendix F, enforcing a margin Γ that strictly exceeds the model's inherent semantic separation capacity $C_{\mathcal{H}}$ causes the approximation gap $\mathcal{A}_{\Gamma}(\mathcal{H})$ to saturate, rendering the consistency bounds vacuous.

If a uniform global margin were applied irrespective of semantic density (as with standard PPO's fixed ϵ), say $\Gamma \equiv \epsilon_{\max}$, then in semantically dense clusters (where actions a and a' are synonyms), the true semantic capacity approaches zero ($C_{\mathcal{H}} \to 0$). Enforcing $\epsilon_{\max}$ in these regions causes the penalty $\mathcal{A}_{\epsilon_{\max}}(\mathcal{H})$ to saturate, breaking theoretical consistency and inducing the entropy collapse empirically observed in Section 5.

In contrast, by dynamically shrinking $\Gamma(s_t, a_t) \to \epsilon_{\min}$ as the expected semantic distance $D_{\pi_{\text{old}}} \to 0$ (in dense clusters), SAMPPO optimizes the Margin-Capacity Profile by ensuring that the margin requirement relaxes proportionally on indistinguishable actions. By dynamically reducing Γ toward $\epsilon_{\min}$ as the local capacity $C_{\mathcal{H}} \to 0$ in dense synonym clusters, SAMPPO actively minimizes the extent to which Γ exceeds the collapsing local capacity, keeping the capacity penalty $\mathcal{A}_{\Gamma}(\mathcal{H})$ as small as feasible and maintaining the consistency bound as tight as possible.

D.7 Discussion

The key implication of this analysis is that SAMPPO can be understood as optimizing a statistically consistent surrogate objective under a non-uniform geometric constraint. By reweighting policy divergence according to semantic structure, the algorithm prioritizes stability in regions of high ambiguity while preserving learning efficiency elsewhere.

This perspective connects trust-region reinforcement learning with structured prediction and $\mathcal{H}$-consistency, where loss functions are routinely adapted to reflect task-specific geometry. It also provides a principled explanation for the empirical stability gains observed in Section 5: by enforcing a structure-aware bound and optimizing the margin-capacity profile dynamically, we prevent the unaligned updates that lead to entropy collapse in standard PPO, ensuring convergence to a robust stationary point.

E Omitted Proofs

E.1 Proof of Theorem 3

Theorem 3 (Entropy Collapse under Mean-Centered Noisy Advantage)**.** *Under the described setup, assuming a small-cluster regime with $(\kappa - 1)\epsilon < 1$, the PPO update produces a policy with entropy:*

$$H(\pi_{new}(\cdot \mid s)) = \log \kappa - \frac{(\kappa-1)\epsilon^2}{2} + \frac{(\kappa-1)(\kappa-2)\epsilon^3}{6} + \mathcal{O}(\epsilon^4). \tag{3}$$

Thus, the entropy drop is strictly positive for $\epsilon > 0$, approximated by: $H(\pi_{old}(\cdot \mid s)) - H(\pi_{new}(\cdot \mid s)) \approx \frac{\kappa-1}{2}\epsilon^2$. For larger clusters outside the convergence radius of this expansion ($\kappa\epsilon \gg 1$), the exact entropy drop scales logarithmically as $\mathcal{O}(\epsilon \log \kappa)$. In both regimes, the entropy drop scales monotonically *with cluster size κ: semantically denser regions suffer greater entropy collapse.*

Proof. The initial policy π_{old} is uniform with entropy $H(\pi_{\text{old}}) = \log \kappa$. The updated policy has $\pi_{\text{new}}(a_1|s) = (1 + (\kappa-1)\epsilon)/\kappa$ and $\pi_{\text{new}}(a_j|s) = (1-\epsilon)/\kappa$ for $j \geq 2$, which sum to 1 by construction.

The new entropy is $H(\pi_{\text{new}}) = -\sum_{i=1}^{\kappa} \pi_{\text{new}}(a_i|s) \log \pi_{\text{new}}(a_i|s)$. Substituting the probabilities and separating the $\log \kappa$ term (using $\sum_i \pi_{\text{new}}(a_i|s) = 1$):

$$H(\pi_{\text{new}}) = \log \kappa - \frac{1}{\kappa} \underbrace{\left[(1 + (\kappa-1)\epsilon) \log(1 + (\kappa-1)\epsilon) + (\kappa-1)(1-\epsilon)\log(1-\epsilon)\right]}_{:=T}.$$

For the regime where $(\kappa - 1)\epsilon < 1$, we use the auxiliary function $f(x) = (1+x)\log(1+x)$ with third-order Maclaurin expansion $f(x) = x + \frac{x^2}{2} - \frac{x^3}{6} + \mathcal{O}(x^4)$, so that $T = f((\kappa-1)\epsilon) + (\kappa-1)f(-\epsilon)$.

Summing the components (linear terms cancel exactly):

$$T = \frac{\epsilon^2}{2}\left[(\kappa-1)^2 + (\kappa-1)\right] + \frac{\epsilon^3}{6}\left[-(\kappa-1)^3 + (\kappa-1)\right] + \mathcal{O}(\epsilon^4)$$
$$= \frac{\kappa(\kappa-1)\epsilon^2}{2} + \frac{(\kappa-1)\epsilon^3}{6}\underbrace{\left[1-(\kappa-1)^2\right]}_{-\kappa(\kappa-2)} + \mathcal{O}(\epsilon^4) = \frac{\kappa(\kappa-1)\epsilon^2}{2} - \frac{\kappa(\kappa-1)(\kappa-2)\epsilon^3}{6} + \mathcal{O}(\epsilon^4).$$

Dividing by κ and subtracting from $\log\kappa$ yields the stated entropy drop, whose leading term $\frac{\kappa-1}{2}\epsilon^2$ is strictly positive.

For massive clusters outside the convergence radius of this expansion (i.e., when $\epsilon\kappa \gg 1$), the linear Taylor expansion is no longer valid. Instead, we evaluate the asymptotic behavior of T directly. For large κ, the terms in T can be approximated by keeping the dominant components. Approximating $1 + (\kappa-1)\epsilon \approx \kappa\epsilon$ and $\kappa - 1 \approx \kappa$, we have:

$$T \approx \kappa\epsilon\log(\kappa\epsilon) + \kappa(1-\epsilon)\log(1-\epsilon).$$

Dividing by κ, the normalized term evaluates to:

$$\frac{1}{\kappa}T \approx \epsilon\log(\kappa\epsilon) + (1-\epsilon)\log(1-\epsilon) = \epsilon\log\kappa + \epsilon\log\epsilon + (1-\epsilon)\log(1-\epsilon).$$

Since $H(\pi_{\text{new}}) = \log\kappa - \frac{1}{\kappa}T$, the exact entropy drop evaluates to:

$$H(\pi_{\text{old}}) - H(\pi_{\text{new}}) = \frac{1}{\kappa}T \approx \epsilon\log\kappa + \epsilon\log\epsilon + (1-\epsilon)\log(1-\epsilon).$$

As κ grows large, the positive term $\epsilon\log\kappa$ dominates. Since the terms depending solely on ϵ are bounded constants, the entropy drop scales logarithmically as $\mathcal{O}(\epsilon\log\kappa)$. To verify monotonicity formally, note that the exact (non-asymptotic) entropy drop is $\frac{1}{\kappa}T = \frac{1+(\kappa-1)\epsilon}{\kappa}\log\left(\frac{1+(\kappa-1)\epsilon}{\kappa}\right) + \frac{(\kappa-1)(1-\epsilon)}{\kappa}\log\left(\frac{1-\epsilon}{\kappa}\right) + \log\kappa = \mathsf{D}_{\mathrm{KL}}(\pi_{\text{new}} \,\|\, \pi_{\text{old}}) \geq 0$. Now write $T(\kappa) = f\big((\kappa-1)\epsilon\big) + (\kappa-1)f(-\epsilon)$ with $f(x) = (1+x)\log(1+x)$ as above. Since f is convex, the first term is convex in κ (a convex function composed with an affine map) and the second is linear in κ, so T is convex in κ with $T(1) = 0$. For a convex function vanishing at $\kappa = 1$, the secant slope $T(\kappa)/(\kappa-1)$ is non-decreasing in κ. Hence $T(\kappa)/\kappa = \frac{T(\kappa)}{\kappa-1}\cdot\frac{\kappa-1}{\kappa}$ is a product of two non-negative non-decreasing functions of κ and is therefore non-decreasing (strictly increasing for $\epsilon > 0$). In both regimes, the exact entropy drop $H(\pi_{\text{old}}) - H(\pi_{\text{new}}) = T(\kappa)/\kappa$ thus grows *monotonically* with cluster size κ, completing the proof. □

Interpretation. The theorem reveals a qualitatively strong and logically consistent result. Whether scaling linearly as $\mathcal{O}(\kappa\epsilon^2)$ in small clusters or logarithmically $\mathcal{O}(\epsilon\log\kappa)$ for massive clusters, the implications are the same:

Dense regions are most susceptible: The correct centered formulation shows that larger synonym clusters suffer systematically larger entropy collapse.

Coupling through the simplex: The collapse is driven not by the gradient on a_1, but by all $\kappa - 1$ synonyms simultaneously hitting their lower clip bounds, with the sum-to-one constraint forcing $\pi_{\text{new}}(a_1|s)$ beyond the upper clip.

Trust-Region Amplification: Larger ϵ and larger κ *compound* the collapse, providing the direct justification for SAMPPO's adaptive tightening of the trust region in semantically dense neighborhoods.

E.2 Proof of Corollary 4

Corollary 4 (Expected Entropy Decay under Exchangeable Noise)**.** *Under the setup of Theorem 3, suppose the rewarded synonym is drawn uniformly at random from the cluster (exchangeable noise), so that* $\mathbb{E}[\pi_{new}(\cdot|s)] = \pi_{old}(\cdot|s)$*. Then, by symmetry, the one-step entropy drop is the same deterministic quantity for every realization:* $\mathbb{E}[H(\pi_{new})] = H(\pi_{old}) - \frac{\kappa-1}{2}\epsilon^2 + \mathcal{O}(\epsilon^3)$*. More generally, for* any *unbiased stochastic update with* $\mathbb{E}[\pi_{new}] = \pi_{old}$ *uniform on the cluster, Jensen's inequality applied to the strictly concave entropy functional yields* $\mathbb{E}[H(\pi_{new})] \leq H(\mathbb{E}[\pi_{new}]) = H(\pi_{old})$*, with strict inequality unless the update is degenerate.*

Proof. (i) Exchangeable single-winner noise. Suppose the index I of the rewarded synonym is drawn uniformly from $\{1, \dots, \kappa\}$, independently of the state. Conditioned on $I = i$, the setup of Theorem 3 applies verbatim with a_i in the role of a_1: the update produces $\pi_{\text{new}}(a_i|s) = (1 + (\kappa - 1)\epsilon)/\kappa$ and $\pi_{\text{new}}(a_j|s) = (1 - \epsilon)/\kappa$ for $j \neq i$. Since entropy is invariant under permutations of the probability vector, the resulting entropy is the *same deterministic quantity* for every realization i, namely the value computed in Theorem 3:

$$H(\pi_{\text{new}}(\cdot|s)) = \log \kappa - \frac{(\kappa - 1)\epsilon^2}{2} + \mathcal{O}(\epsilon^3).$$

Taking the expectation over I leaves this value unchanged, which proves the first claim. Note that $\mathbb{E}[\pi_{\text{new}}(a_j|s)] = \frac{1}{\kappa} \cdot \frac{1+(\kappa-1)\epsilon}{\kappa} + \frac{\kappa-1}{\kappa} \cdot \frac{1-\epsilon}{\kappa} = \frac{1}{\kappa}$ for every j: the *expected policy* is exactly the uniform reference policy, yet the expected entropy strictly drops.

(ii) General unbiased updates. Let π_{new} be any random updated policy with $\mathbb{E}[\pi_{\text{new}}] = \pi_{\text{old}}$ uniform on the cluster. The entropy functional H is strictly concave on the simplex, so by Jensen's inequality,

$$\mathbb{E}[H(\pi_{\text{new}})] \leq H(\mathbb{E}[\pi_{\text{new}}]) = H(\pi_{\text{old}}),$$

with equality if and only if π_{new} is almost surely constant, i.e., the update is degenerate. Hence symmetry of the expected update cannot prevent systematic entropy loss: any non-degenerate zero-mean stochastic oscillation strictly decreases expected entropy. □

E.3 Proof of Lemma 6

Lemma 6 (Equivalence to Average KL)**.** *Let the state-level isolation score $D(s)$ be bounded such that $0 \leq D_{\min} \leq D(s) \leq D_{\max}$. Let the weight function be $w(s) = \frac{1}{D(s)+\nu}$ for some constant $\nu > 0$. The structure-aware divergence $D_w(\pi, \pi_{old})$ and the standard average KL divergence $\mathsf{D}_{\text{KL}}^{avg}$ satisfy the following two-sided inequality:*

$$(D_{\min} + \nu) \cdot D_w(\pi, \pi_{old}) \leq \mathsf{D}_{\text{KL}}^{avg} \leq (D_{\max} + \nu) \cdot D_w(\pi, \pi_{old}). \tag{6}$$

Proof. We begin with the definition of the average KL divergence and introduce the weight function $w(s)$ and its inverse, $1/w(s) = D(s) + \nu$.

$$\begin{aligned} \mathsf{D}_{\text{KL}}^{\text{avg}} &= \mathop{\mathbb{E}}_{s\sim\rho_{\pi_{\text{old}}}} \left[\frac{1}{w(s)} \cdot w(s) \cdot \mathsf{D}_{\text{KL}}(\pi(\cdot|s) \parallel \pi_{\text{old}}(\cdot|s)) \right] \\ &= \mathop{\mathbb{E}}_{s\sim\rho_{\pi_{\text{old}}}} \left[(D(s) + \nu) \cdot w(s) \cdot \mathsf{D}_{\text{KL}}(\pi(\cdot|s) \parallel \pi_{\text{old}}(\cdot|s)) \right]. \end{aligned}$$

By assumption, we have the lower bound $D(s) \geq D_{\min}$ for any state s. Therefore, $D(s)+\nu \geq D_{\min}+\nu$. Applying this to the non-negative expectation yields $\mathsf{D}_{\text{KL}}^{\text{avg}} \geq (D_{\min} + \nu) \cdot D_w(\pi, \pi_{\text{old}})$. The upper bound follows symmetrically using $D(s) \leq D_{\max}$, completing the proof. □

E.4 Proof of Theorem 7

Theorem 7 (Structure-Aware Surrogate Improvement)**.** *Let $C_0 = \frac{2\gamma}{(1-\gamma)^2} \max_{s,a} |A_{\pi_{old}}(s, a)|$ and $C = C_0 \cdot (D_{\max} + \nu)$, where $J(\pi)$ denotes the expected return.*

(a) ***Rigorous bound.*** *The structure-aware surrogate objective satisfies:*

$$J(\pi) \geq J(\pi_{old}) + L_{\pi_{old}}(\pi) - C \cdot \max_s \big[w(s) \mathsf{D}_{\text{KL}}\big(\pi(\cdot|s) \parallel \pi_{old}(\cdot|s)\big)\big]. \tag{7}$$

(b) ***Practical surrogate.*** *As in Schulman et al. [2015], replacing the intractable worst-case maximum with the expected value under $\rho_{\pi_{old}}$ yields the surrogate used in practice:*

$$\widetilde{J}(\pi) = J(\pi_{old}) + L_{\pi_{old}}(\pi) - C \cdot D_w(\pi, \pi_{old}). \tag{8}$$

Since $\mathbb{E}_s[\cdot] \leq \max_s[\cdot]$, the surrogate $\widetilde{J}(\pi)$ upper-bounds the rigorous bound (7); it is therefore a less conservative (but still consistent) proxy for policy improvement.

(c) ***Rigorous average-divergence bound.*** *Let $\zeta^\pi = \max_s \big|\mathbb{E}_{a\sim\pi(\cdot|s)}[A_{\pi_{old}}(s, a)]\big|$, so that $\frac{2\gamma\zeta^\pi}{(1-\gamma)^2} \leq C_0$. Then*

$$J(\pi) \geq J(\pi_{old}) + L_{\pi_{old}}(\pi) - \frac{2\gamma\zeta^\pi}{(1-\gamma)^2} \sqrt{\frac{D_{\max} + \nu}{2} D_w(\pi, \pi_{old})}. \tag{9}$$

Proof. From standard trust-region theory [Schulman et al., 2015], the theoretical expected return difference is formally bounded by the maximum total variation (TV) distance. Since the TV distance is symmetric, Pinsker's inequality ($D_{\text{TV}}^2 \le \frac{1}{2}\mathsf{D}_{\text{KL}}$) bounds the square of the TV distance by the maximum KL divergence $\max_s \mathsf{D}_{\text{KL}}(\pi(\cdot|s) \,\|\, \pi_{\text{old}}(\cdot|s))$ regardless of the argument order. Thus:

$$J(\pi) \ge J(\pi_{\text{old}}) + L_{\pi_{\text{old}}}(\pi) - C_0 \max_s \mathsf{D}_{\text{KL}}(\pi(\cdot|s) \,\|\, \pi_{\text{old}}(\cdot|s)),$$

where $C_0 = \frac{2\gamma}{(1-\gamma)^2} \max_{s,a} |A_{\pi_{\text{old}}}(s,a)|$.

To incorporate our structure-aware divergence, we observe that for any state s, the weight satisfies $w(s) = \frac{1}{D(s)+\nu} \ge \frac{1}{D_{\max}+\nu}$. Rearranging this gives the pointwise bound:

$$\mathsf{D}_{\text{KL}}(\pi(\cdot|s) \,\|\, \pi_{\text{old}}(\cdot|s)) \le (D_{\max} + \nu) \cdot w(s) \mathsf{D}_{\text{KL}}(\pi(\cdot|s) \,\|\, \pi_{\text{old}}(\cdot|s)).$$

Taking the maximum over all states yields:

$$\max_s \mathsf{D}_{\text{KL}}(\cdot \,\|\, \cdot) \le (D_{\max} + \nu) \max_s [w(s) \mathsf{D}_{\text{KL}}(\cdot \,\|\, \cdot)].$$

Substituting into the standard trust-region bound and defining $C = C_0(D_{\max} + \nu)$ establishes part (a).

For part (b), note that for any non-negative function f, $\mathbb{E}_s[f(s)] \le \max_s f(s)$, so replacing the maximum in (7) with the expectation $D_w(\pi, \pi_{\text{old}})$ yields a surrogate that is at least as large as the rigorous lower bound. This is the same relaxation employed by Schulman et al. [2015] to obtain a tractable optimization objective: it trades the intractable worst-case penalty for a tractable average-case proxy, producing a less conservative but practically effective optimization target.

For part (c), we invoke the performance-difference bound of Achiam et al. [2017, Corollary 1]: for any two policies,

$$J(\pi) - J(\pi_{\text{old}}) \ge L_{\pi_{\text{old}}}(\pi) - \frac{2\gamma\zeta^\pi}{(1-\gamma)^2} \mathop{\mathbb{E}}_{s\sim\rho_{\pi_{\text{old}}}} \big[D_{\text{TV}}\big(\pi(\cdot|s), \pi_{\text{old}}(\cdot|s)\big)\big],$$

where $\zeta^\pi = \max_s \big|\mathbb{E}_{a\sim\pi(\cdot|s)}[A_{\pi_{\text{old}}}(s,a)]\big|$, D_{TV} denotes the total variation distance, and we identify the normalized discounted state-visitation distribution of π_{old} with $\rho_{\pi_{\text{old}}}$, absorbing normalization constants into $L_{\pi_{\text{old}}}(\pi)$ as in part (a). Crucially, in contrast to the worst-case bound underlying part (a), the divergence penalty here is an *expectation* over the state distribution. By Pinsker's inequality, $D_{\text{TV}} \le \sqrt{\mathsf{D}_{\text{KL}}/2}$ pointwise, and introducing the structure-aware weight as in part (a), $\mathsf{D}_{\text{KL}} = (D(s) + \nu)\, w(s)\, \mathsf{D}_{\text{KL}} \le (D_{\max} + \nu)\, w(s)\, \mathsf{D}_{\text{KL}}$, so that

$$\mathop{\mathbb{E}}_s[D_{\text{TV}}] \le \mathop{\mathbb{E}}_s\left[\sqrt{\tfrac{1}{2}(D_{\max}+\nu) w(s) \mathsf{D}_{\text{KL}}}\right] \le \sqrt{\tfrac{1}{2}(D_{\max}+\nu) \mathop{\mathbb{E}}_s[w(s)\mathsf{D}_{\text{KL}}]} = \sqrt{\tfrac{D_{\max}+\nu}{2} D_w(\pi, \pi_{\text{old}})},$$

where the second inequality follows from Jensen's inequality applied to the concave square root. Substituting this bound into the inequality above yields part (c). Finally, since $\zeta^\pi \le \max_{s,a}|A_{\pi_{\text{old}}}(s,a)|$, the leading constant satisfies $\frac{2\gamma\zeta^\pi}{(1-\gamma)^2} \le C_0$: the rigorous average-divergence bound holds with a constant no larger than that of part (a). □

E.5 Proof of Proposition 8 and Supporting Remarks

Proposition 8 (Adaptive Clipping Realizes a Uniform Structure-Aware KL Budget)**.** *For each state s, let $\epsilon(s) = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot \min(D(s)/D_{bound}, 1)$ denote the clip radius evaluated at the state-level isolation score $D(s)$, and suppose the updated policy satisfies the idealized clip condition $|\pi(a|s)/\pi_{old}(a|s) - 1| \le \epsilon(s)$ for all a, and assume $\epsilon_{\max} < 1$. Then*

$$\mathsf{D}_{\text{KL}}(\pi(\cdot|s) \,\|\, \pi_{old}(\cdot|s)) \le \frac{\epsilon(s)^2}{2(1-\epsilon_{\max})},$$

and therefore, with $w(s) = 1/(D(s)+\nu)$:

$$w(s)\mathsf{D}_{\text{KL}}(\pi(\cdot|s) \,\|\, \pi_{old}(\cdot|s)) \le \delta^* := \sup_s \frac{w(s)\epsilon(s)^2}{2(1-\epsilon_{\max})} \le \frac{\epsilon_{\min}^2/\nu + (\epsilon_{\max}^2 - \epsilon_{\min}^2)/D_{bound}}{2(1-\epsilon_{\max})} < +\infty,$$

uniformly over all states.

Proof. Write $r(a) = \pi(a|s)/\pi_{\text{old}}(a|s)$ and $\varphi(x) = x \log x - x + 1$. Since $\sum_a \pi_{\text{old}}(a)\,(r(a) - 1) = 0$, the per-state KL divergence can be written as

$$\mathsf{D}_{\mathrm{KL}}(\pi(\cdot|s) \,\|\, \pi_{\text{old}}(\cdot|s)) = \sum_a \pi_{\text{old}}(a)\,\varphi(r(a)).$$

As $\varphi(1) = \varphi'(1) = 0$ and $\varphi''(x) = 1/x \le 1/(1 - \epsilon_{\max})$ on the clip interval $[1 - \epsilon(s), 1 + \epsilon(s)]$, Taylor's theorem gives

$$\varphi(r(a)) \le \frac{(r(a) - 1)^2}{2(1 - \epsilon_{\max})} \le \frac{\epsilon(s)^2}{2(1 - \epsilon_{\max})};$$

averaging over π_{old} yields the first inequality. Multiplying by $w(s)$, it remains to bound $\sup_{D \ge 0} \epsilon(D)^2/(D + \nu)$. For $D \le D_{\text{bound}}$, writing $m = D/D_{\text{bound}} \le 1$, we have

$$\epsilon(D)^2 = \epsilon_{\min}^2 + (\epsilon_{\max} - \epsilon_{\min})\, m(2\epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min})m) \le \epsilon_{\min}^2 + \big(\epsilon_{\max}^2 - \epsilon_{\min}^2\big)m,$$

and since $m/(D+\nu) \le 1/D_{\text{bound}}$, it follows that $\epsilon(D)^2/(D+\nu) \le \epsilon_{\min}^2/\nu + (\epsilon_{\max}^2 - \epsilon_{\min}^2)/D_{\text{bound}}$. For $D > D_{\text{bound}}$, splitting $\epsilon_{\max}^2 = \epsilon_{\min}^2 + (\epsilon_{\max}^2 - \epsilon_{\min}^2)$ gives $\epsilon_{\max}^2/(D+\nu) \le \epsilon_{\min}^2/\nu + (\epsilon_{\max}^2 - \epsilon_{\min}^2)/D_{\text{bound}}$ as well. Dividing by $2(1 - \epsilon_{\max})$ yields δ^*. □

Remark 12 (Per-action vs. per-state schedules)**.** *The implementation evaluates ϵ_t at the sampled action's score $D_{\pi_{old}}(s_t, a_t)$ rather than the state-level $D(s) = \mathbb{E}_a[D(s,a)]$. To verify that the global budget δ^* is respected, one must bound $\mathbb{E}_a[\epsilon(D(s,a))^2]$. Although $\epsilon(\cdot)$ itself is concave, its square restricted to the affine region is convex, so Jensen's inequality cannot be applied directly to $\epsilon(\cdot)^2$. Instead, the proof of Proposition 8 already upper-bounds $\epsilon(D)^2$ by the concave envelope $h(D) = \epsilon_{\min}^2 + (\epsilon_{\max}^2 - \epsilon_{\min}^2)\min(D/D_{bound}, 1)$. Since h is concave, Jensen's inequality gives $\mathbb{E}_a[\epsilon(D(s,a))^2] \le \mathbb{E}_a[h(D(s,a))] \le h(D(s))$, and $h(D(s))$ is the exact expression used to establish δ^*. Thus, the per-action schedule respects the budget δ^* in expectation.*

Remark 13 (Relation to clipping and duality)**.** *As in standard PPO's relationship to TRPO, clipping enforces the ratio condition at the level of gradient incentives rather than as a hard constraint (indeed, our Theorem 3 shows the simplex constraint can push a realized ratio past the clip). The proposition therefore characterizes the trust region that SAMPPO's surrogate is constructed to approximate, precisely the sense in which uniform PPO approximates a uniform KL trust region. In the corresponding penalized (dual) form, a state-dependent multiplier $\lambda(s) = w(s) = 1/(D(s) + \nu)$ recovers the density-weighted penalty $\mathbb{E}_s[w(s)\mathsf{D}_{\mathrm{KL}}] = D_w$; note that $\epsilon_{\min} > 0$ plays precisely the role of the stability constant ν, keeping the implicit penalty bounded in maximally dense regions.*

Thus, SAMPPO's action-dependent clipping realizes the structure-aware mirror descent geometry of D_w with a provable uniform budget δ^*, while avoiding the intractable $\mathcal{O}(|\mathcal{V}|)$ overhead of computing full-vocabulary expectations. The surrogate approximation error is separately quantified in Proposition 14.

E.5.1 Empirical Verification of the KL Budget

Because clipped surrogates do not hard-enforce the idealized clip condition, we verified the budget of Proposition 8 directly on our Llama-3.1-8B HH-RLHF runs. At every optimization iteration, we measured the realized per-update KL divergence $\mathsf{D}_{\mathrm{KL}}(\pi_\theta(\cdot|s) \,\|\, \pi_{\text{old}}(\cdot|s))$ between consecutive policies for each logged state s (not the cumulative divergence from the initial policy reported in the reward-vs-KL frontier of Figure 3), and weighted it by $w(s) = 1/(D(s) + \nu)$. With $\nu = 0.05$, on the order of the smallest observed state-level isolation scores, the resulting budget evaluates to $\delta^* = 0.33$. Across all logged optimization steps, 99.6% of states satisfy $w(s)\mathsf{D}_{\mathrm{KL}} \le \delta^*$; the rare exceedances are small (median $+8\%$ above δ^*) and are consistent with the simplex-coupling effect predicted by Theorem 3, in which the sum-to-one constraint pushes a rewarded action's realized ratio slightly past its clip bound. Conclusions are unchanged for $\nu \in \{0.02, 0.1\}$. The idealized budget is thus respected in practice.

E.6 Proof of Lemma 11

Lemma 11 (Structure-Aware Bregman Divergence)**.** *Let the step-dependent structure-aware divergence for mirror descent be defined via the KL divergence: $D_{w_k}(\pi, \pi_k) =$*

$\mathbb{E}_{s\sim\rho_{\pi_k}}[w_k(s)\mathsf{D}_{\mathrm{KL}}(\pi(\cdot|s)\,\|\,\pi_k(\cdot|s))]$. *This is a valid Bregman divergence generated by the state-weighted negative entropy:*

$$\Psi_{w_k}(\pi) = \mathop{\mathbb{E}}_{s\sim\rho_{\pi_k}}\left[w_k(s)\sum_a \pi(a|s)\log\pi(a|s)\right]. \tag{19}$$

Furthermore, D_{w_k} is strongly convex with respect to the expected squared ℓ_1-norm ($\mathbb{E}_s[\|\pi(\cdot|s) - \pi_k(\cdot|s)\|_1^2]$) with a uniform global modulus $\mu = \frac{1}{D_{\max}+\nu} > 0$.

Proof. The Bregman divergence generated by Ψ_{w_k} is defined as:

$$B_{\Psi_{w_k}}(\pi, \pi_k) = \Psi_{w_k}(\pi) - \Psi_{w_k}(\pi_k) - \langle\nabla\Psi_{w_k}(\pi_k), \pi - \pi_k\rangle.$$

Since the state-wise weighting $w_k(s)$ is strictly positive and independent of the optimization variable π, $B_{\Psi_{w_k}}$ evaluates exactly to the weighted sum of state-wise Bregman divergences of the unweighted negative entropy, which mathematically yields the KL divergence $\mathsf{D}_{\mathrm{KL}}(\pi\|\pi_k)$. Thus, $B_{\Psi_{w_k}}(\pi, \pi_k) = \mathbb{E}_{s\sim\rho_{\pi_k}}[w_k(s)\mathsf{D}_{\mathrm{KL}}(\pi(\cdot|s)\,\|\,\pi_k(\cdot|s))] = D_{w_k}(\pi, \pi_k)$.

By Pinsker's inequality, $\mathsf{D}_{\mathrm{KL}}(p\,\|\,q) \geq \frac{1}{2}\|p - q\|_1^2$. Since $w_k(s) = \frac{1}{D_{\pi_k}(s)+\nu} \geq \frac{1}{D_{\max}+\nu}$, we have:

$$D_{w_k}(\pi, \pi_k) \geq \frac{1}{2(D_{\max}+\nu)}\mathop{\mathbb{E}}_s\left[\|\pi(\cdot|s) - \pi_k(\cdot|s)\|_1^2\right],$$

establishing strong convexity with a uniform global bound μ. □

E.7 Proof of Theorem 9

Theorem 9 (Convergence to a Stationary Point)**.** *Assume the expected policy objective $J(\pi)$ is L-smooth uniformly across shifting state distributions ρ_{π_k} with respect to the expected squared ℓ_1-norm ($\mathbb{E}_s[\|\pi(\cdot|s) - \pi_k(\cdot|s)\|_1^2]$), implicitly absorbing a worst-case concentrability coefficient standard in MDP analysis, and bounded above by J^*. Let the sequence of policies $\{\pi_k\}_{k=0}^T$ be generated by the SAMPPO mirror descent update:*

$$\pi_{k+1} = \mathop{\mathrm{argmax}}_{\pi}\left\{\langle\nabla J(\pi_k), \pi\rangle - \frac{1}{\eta}D_{w_k}(\pi, \pi_k)\right\}. \tag{10}$$

For a step size $\eta \leq \frac{\mu}{L}$, SAMPPO converges to a stationary point at a rate of $\mathcal{O}(1/T)$:

$$\min_{0\leq k<T}\mathop{\mathbb{E}}_s\left[\|\pi_{k+1} - \pi_k\|_1^2\right] \leq \frac{2\eta(D_{\max}+\nu)}{T}(J^* - J(\pi_0)). \tag{11}$$

Proof. By the standard descent lemma for L-smooth functions (when maximizing J), we have:

$$J(\pi_{k+1}) - J(\pi_k) \geq \langle\nabla J(\pi_k), \pi_{k+1} - \pi_k\rangle - \frac{L}{2}\mathop{\mathbb{E}}_s\left[\|\pi_{k+1} - \pi_k\|_1^2\right].$$

The first-order optimality condition of the mirror descent update yields:

$$\left\langle\nabla J(\pi_k) - \frac{1}{\eta}\nabla_\pi D_{w_k}(\pi, \pi_k)\Big|_{\pi=\pi_{k+1}},\ \pi - \pi_{k+1}\right\rangle \leq 0 \quad \text{for all valid } \pi.$$

Recalling that $\nabla_\pi D_{w_k}(\pi, \pi_k) = \nabla\Psi_{w_k}(\pi) - \nabla\Psi_{w_k}(\pi_k)$, we set the test policy to $\pi = \pi_k$ to obtain:

$$\langle\nabla J(\pi_k), \pi_{k+1} - \pi_k\rangle \geq \frac{1}{\eta}\langle\nabla\Psi_{w_k}(\pi_{k+1}) - \nabla\Psi_{w_k}(\pi_k), \pi_{k+1} - \pi_k\rangle.$$

A fundamental property of Bregman divergences is that the inner product of the gradient differences with the point differences evaluates exactly to the symmetrized Bregman divergence:

$$\langle\nabla\Psi_{w_k}(x) - \nabla\Psi_{w_k}(y), x - y\rangle = B_{\Psi_{w_k}}(x, y) + B_{\Psi_{w_k}}(y, x).$$

Applying this to our inner product:

$$\langle\nabla J(\pi_k), \pi_{k+1} - \pi_k\rangle \geq \frac{1}{\eta}\Big(D_{w_k}(\pi_{k+1}, \pi_k) + D_{w_k}(\pi_k, \pi_{k+1})\Big).$$

While the weight function $w_k(s)$ is step-dependent, Lemma 11 guarantees the strong convexity modulus $\mu = \frac{1}{D_{\max}+\nu}$ provides a strict, uniform global lower bound across all iterations ($w_k(s) \geq \mu$). Applying this uniform μ-strong convexity to both terms yields:

$$\langle \nabla J(\pi_k), \pi_{k+1} - \pi_k \rangle \geq \frac{1}{\eta} \left(\frac{\mu}{2} \mathop{\mathbb{E}}_{s} \left[\|\pi_{k+1} - \pi_k\|_1^2 \right] + \frac{\mu}{2} \mathop{\mathbb{E}}_{s} \left[\|\pi_k - \pi_{k+1}\|_1^2 \right] \right) = \frac{\mu}{\eta} \mathop{\mathbb{E}}_{s} \left[\|\pi_{k+1} - \pi_k\|_1^2 \right].$$

Substituting this exact bound into the descent lemma provides:

$$J(\pi_{k+1}) - J(\pi_k) \geq \left(\frac{\mu}{\eta} - \frac{L}{2} \right) \mathop{\mathbb{E}}_{s} \left[\|\pi_{k+1} - \pi_k\|_1^2 \right].$$

Setting the step size $\eta \leq \mu/L$ ensures $\frac{\mu}{\eta} \geq L$, which mathematically guarantees $\frac{\mu}{\eta} - \frac{L}{2} \geq \frac{\mu}{\eta} - \frac{\mu}{2\eta} = \frac{\mu}{2\eta}$. The coefficient is therefore strictly positive, giving:

$$J(\pi_{k+1}) - J(\pi_k) \geq \frac{\mu}{2\eta} \mathop{\mathbb{E}}_{s} \left[\|\pi_{k+1} - \pi_k\|_1^2 \right].$$

Summing this telescoping inequality over T iterations yields:

$$J(\pi_T) - J(\pi_0) \geq \frac{\mu}{2\eta} \sum_{k=0}^{T-1} \mathop{\mathbb{E}}_{s} \left[\|\pi_{k+1} - \pi_k\|_1^2 \right] \geq \frac{\mu T}{2\eta} \min_{0 \leq k < T} \mathop{\mathbb{E}}_{s} \left[\|\pi_{k+1} - \pi_k\|_1^2 \right].$$

Bounding $J(\pi_T) \leq J^*$, rearranging to isolate the minimum policy update, and substituting $\mu = \frac{1}{D_{\max}+\nu}$ directly yields the stated bound:

$$\min_{0 \leq k < T} \mathop{\mathbb{E}}_{s} \left[\|\pi_{k+1} - \pi_k\|_1^2 \right] \leq \frac{2\eta}{\mu T} (J^* - J(\pi_0)) = \frac{2\eta(D_{\max} + \nu)}{T} (J^* - J(\pi_0)).$$

This completes the proof. □

E.8 Proof of Theorem 10

Theorem 10 (Implicit Structure-Aware $\mathcal{H}$-Consistency)**.** *Let $h_\theta(s,a) = \log \pi_\theta(a|s) - \log \pi_{\text{old}}(a|s)$ be the relative scoring function of the policy hypothesis set $\mathcal{H}$. Up to a second-order approximation error uniformly bounded within the trust region (as explicitly quantified in Proposition 14), the SAMPPO clipped objective is mathematically equivalent to minimizing an advantage-weighted structure-aware margin-shifted Hinge surrogate:*

$$\mathsf{L}_{\Phi,\Gamma}(h_\theta) = |A_t| \Phi_{\text{hinge}}(y_t \cdot h_\theta(s_t,a_t) - \Gamma(s_t,a_t)), \tag{12}$$

where $y_t = \text{sign}(A_t)$, $\Phi_{\text{hinge}}(u) = \max(0,-u)$, and the local margin is $\Gamma(s_t,a_t) = \epsilon_t$. Thus, SAMPPO satisfies the established structure-aware $\mathcal{H}$-consistency bound:

$$\mathcal{R}(h_\theta) - \mathcal{R}^*(\mathcal{H}) + \mathcal{M}(\mathcal{H}) \leq \mathcal{R}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(h_\theta) - \mathcal{R}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H}) + \mathcal{M}_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H}) + \mathcal{A}_\Gamma(\mathcal{H}), \tag{13}$$

where $\mathcal{R}(h_\theta)$ is the target 0-1 ranking error, $\mathcal{M}(\mathcal{H})$ is the minimizability gap, $\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma} = \mathsf{L}_{\Phi_{\text{hinge}},\Gamma}/\Phi_{\text{hinge}}(-\Gamma)$ is the inverse-margin weighted surrogate loss, and $\mathcal{A}_\Gamma(\mathcal{H}) = \mathbb{E}[\mathcal{C}^*_{\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H})] - \mathbb{E}[\mathcal{C}^*(\mathcal{H})]$ is the structure-aware approximation gap.

Proof. The probability ratio in the SAMPPO objective is $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)} = \exp(h_\theta(s_t,a_t))$. Let $y_t = \text{sign}(A_t)$ be the preference label induced by the advantage. We analyze the optimization of the pointwise SAMPPO loss $\ell_t^{\text{SAMPPO}}(h_\theta) = \min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon_t, 1+\epsilon_t)A_t)$ for both positive and negative advantages:

Case 1 ($A_t > 0$, $y_t = 1$)**:** The objective seeks to maximize $A_t \min(r_t(\theta), 1+\epsilon_t)$. Since PPO inherently restricts updates to a small local trust region where $h_\theta \approx 0$, we can apply a first-order Taylor expansion: $\exp(h_\theta) \approx 1 + h_\theta$. The optimization is equivalent to minimizing the negative objective:

$$\ell_+(h_\theta) \approx -A_t \min(1 + h_\theta, 1 + \epsilon_t).$$

Using the algebraic identity $-\min(a,b) = -b + \max(0, b-a)$, we can rewrite this as:

$$\ell_+(h_\theta) \approx -A_t(1+\epsilon_t) + A_t \max(0, \epsilon_t - h_\theta).$$

Since ϵ_t is pre-computed statically via the behavior policy π_{old} and is detached from the computation graph (`stop_gradient`), its gradient with respect to the optimization variable θ is identically zero ($\nabla_\theta \epsilon_t = 0$). Therefore, translating the objective by the constant $-A_t(1+\epsilon_t)$ does not alter the gradient optimization landscape, and minimizing this is mathematically equivalent to minimizing $A_t \max(0, \epsilon_t - h_\theta) = |A_t| \Phi_{\text{hinge}}(y_t \cdot h_\theta - \epsilon_t)$.

Case 2 ($A_t < 0$, $y_t = -1$): The objective seeks to maximize $A_t \max(r_t(\theta), 1-\epsilon_t)$. Using the Taylor expansion, minimizing the negative objective gives:

$$\ell_-(h_\theta) \approx -A_t \max(1+h_\theta, 1-\epsilon_t) = |A_t| \max(1+h_\theta, 1-\epsilon_t).$$

Using the identity $\max(a,b) = b + \max(0, a-b)$, we obtain:

$$\ell_-(h_\theta) \approx |A_t|(1-\epsilon_t) + |A_t| \max(0, h_\theta + \epsilon_t).$$

Ignoring the constant $|A_t|(1-\epsilon_t)$, this minimizes $|A_t| \max(0, \epsilon_t - (-h_\theta)) = |A_t| \Phi_{\text{hinge}}(y_t \cdot h_\theta - \epsilon_t)$.

In both cases, the SAMPPO objective locally minimizes $\mathsf{L}_{\Phi,\Gamma}(h_\theta) = |A_t| \Phi_{\text{hinge}}(y_t \cdot h_\theta(s_t,a_t) - \Gamma(s_t,a_t))$ with a local margin $\Gamma(s_t,a_t) = \epsilon_t$. This precisely matches the definition of a *margin-shifted surrogate*.

To explicitly verify theoretical consistency, we examine the inverse-margin weighted surrogate loss $\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}(u)$ evaluated at $u = y_t \cdot h_\theta$. Retaining the advantage weight $|A_t|$, this normalizes exactly to:

$$|A_t| \widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}(u) = |A_t| \frac{\max(0, \epsilon_t - u)}{\epsilon_t} = |A_t| \max\left(0, 1 - \frac{u}{\epsilon_t}\right).$$

This explicitly recovers the canonical formulation of the margin-scaled hinge loss. Since this normalized surrogate satisfies $\widetilde{\mathsf{L}}_{\Phi_{\text{hinge}},\Gamma}(u) \geq 1$ for all $u \leq 0$, it upper-bounds the advantage-weighted 0-1 classification error ($|A_t| \mathbb{1}_{u \leq 0}$), thereby fulfilling the core mathematical requirement for consistency.

By definition, SAMPPO sets the local margin to $\epsilon_t = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min}) \cdot S_t$. Thus, $\Gamma(s_t,a_t)$ is a *structure-aware margin* that scales proportionally with the semantic isolation of the action space. Therefore, SAMPPO inherently satisfies the Structure-Aware $\mathcal{H}$-Consistency bounds. □

The following proposition quantifies the local approximation in Theorem 10, establishing that the Hinge surrogate equivalence holds with a second-order error uniformly bounded within the trust region.

Proposition 14 (Quantified Local Hinge Approximation)**.** *For each timestep t, define the SAMPPO effective loss as the θ-dependent component of the negated clipped objective:*

$$\ell_t(h_\theta) = \begin{cases} A_t \max\big(0,\ (1+\epsilon_t) - e^{h_\theta(s_t,a_t)}\big) & \text{if } A_t > 0, \\ |A_t| \max\big(0,\ e^{h_\theta(s_t,a_t)} - (1-\epsilon_t)\big) & \text{if } A_t < 0. \end{cases}$$

Then the difference between the exact SAMPPO loss and the advantage-weighted margin-shifted Hinge surrogate of Theorem 10 satisfies:

$$\left|\ell_t(h_\theta) - |A_t| \Phi_{\text{hinge}}(y_t \cdot h_\theta(s_t,a_t) - \epsilon_t)\right| \leq |A_t| \left(e^{h_\theta(s_t,a_t)} - 1 - h_\theta(s_t,a_t)\right). \tag{20}$$

Assuming $\epsilon_{\max} < 1$, within the trust region $|h_\theta(s_t,a_t)| \leq B \coloneqq \log \frac{1}{1-\epsilon_{\max}}$, the right-hand side is further bounded by $\frac{|A_t| h_\theta(s_t,a_t)^2}{2(1-\epsilon_{\max})}$.

Proof. In both cases ($A_t > 0$ and $A_t < 0$), the SAMPPO effective loss and the Hinge surrogate take the form $|A_t| \max(0, f_{y_t}(h))$ and $|A_t| \max(0, g_{y_t}(h))$ respectively, with $h = h_\theta(s_t,a_t)$, where:

$$\begin{aligned} &\text{For } y_t = +1: \quad f(h) = (1+\epsilon_t) - e^h, \quad g(h) = \epsilon_t - h. \\ &\text{For } y_t = -1: \quad f(h) = e^h - (1-\epsilon_t), \quad g(h) = h + \epsilon_t. \end{aligned}$$

In both cases, $f(h) - g(h) = \pm(1 + h - e^h)$, so $|f(h) - g(h)| = e^h - 1 - h$, where the non-negativity follows from the universal inequality $e^x \geq 1 + x$. By the 1-Lipschitz property of $u \mapsto \max(0, u)$:

$$\big|\max(0, f(h)) - \max(0, g(h))\big| \leq |f(h) - g(h)| = e^h - 1 - h.$$

Multiplying by $|A_t|$ yields (20). For the trust region bound, Taylor's theorem with Lagrange remainder gives $e^x - 1 - x = \frac{x^2}{2}e^c$ for some c between 0 and x. Since $|h| \leq B$ implies $e^{|h|} \leq e^B = \frac{1}{1-\epsilon_{\max}}$, we obtain $e^h - 1 - h \leq \frac{h^2}{2(1-\epsilon_{\max})}$. □

This result quantifies the "locally equivalent" characterization of Theorem 10: the equivalence holds up to a second-order error that vanishes quadratically as $h_\theta \to 0$ and is uniformly bounded within the trust region by $\mathcal{O}(\epsilon_{\max}^2)$. For the default configuration $\epsilon_{\max} = 0.30$, the maximum relative approximation error per unit advantage is bounded by $\frac{\log^2(10/7)}{2 \cdot 0.7} < 0.091$, confirming that the Hinge surrogate equivalence is tight within the practical operating regime of the algorithm.

F Supporting Theorems for $\mathcal{H}$-Consistency

To ensure the paper is theoretically self-contained, we formally restate the necessary definitions and bounds for $\mathcal{H}$-consistency evaluated in Theorem 10, as established in the consistency literature [Mohri and Zhong, 2026a].

For an equicontinuous hypothesis set $\mathcal{H}$ bounded by a continuous maximum separation capacity score $C_{\mathcal{H}}$, enforcing a uniform margin Γ restricts the function space, leading to an inherent approximation gap.

Definition 15 (Margin-Capacity Profile)**.** *Let $\mathcal{H}_{C_{\mathcal{H}}}$ be a hypothesis set with maximum score capacity $C_{\mathcal{H}}$. The* Margin-Capacity Profile *of a surrogate Φ is the ratio of the best achievable loss at the capacity limit to the normalization factor:*

$$\rho_\Phi(\Gamma, C_{\mathcal{H}}) = \frac{\Phi(C_{\mathcal{H}} - \Gamma)}{\Phi(-\Gamma)}. \tag{21}$$

For a loss function ℓ and hypothesis set $\mathcal{H}$, let $\mathcal{C}_\ell^*(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{R}_\ell(h)$ denote the best-in-class ℓ-risk (conditional Bayes capacity). The structure-aware approximation gap $\mathcal{A}_\Gamma(\mathcal{H})$ is then defined as the difference in expected attainable capacities between the margin-shifted surrogate loss and the target loss:

$$\mathcal{A}_\Gamma(\mathcal{H}) = \mathbb{E}[\mathcal{C}^*_{\mathsf{L}_{\Phi_{\text{hinge}},\Gamma}}(\mathcal{H})] - \mathbb{E}[\mathcal{C}^*(\mathcal{H})]. \tag{22}$$

If a chosen margin Γ strictly exceeds the available hypothesis capacity $C_{\mathcal{H}}$ ($\Gamma > C_{\mathcal{H}}$), the approximation penalty saturates and becomes vacuous. We formalize this constraint via the following proposition:

Proposition 16 (Penalty under Bounded Capacity)**.** *If the hypothesis set $\mathcal{H}$ has bounded outputs (e.g., due to the activation function used or strict spectral normalization) such that $\sup_{(h,x,y,y') \in \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}} |h(x, y) - h(x, y')| \leq C_{\mathcal{H}}$, then for any margin $\Gamma > C_{\mathcal{H}}$:*

$$\mathcal{A}_\Gamma(\mathcal{H}) \geq \frac{\Phi(C_{\mathcal{H}} - \Gamma)}{\Phi(-\Gamma)} > 0, \tag{23}$$

even if the true ranking error $\mathcal{R}^*(\mathcal{H})$ is zero.

For the margin-shifted Hinge surrogate $\Phi(u) = \max(0, -u)$, if the global margin Γ strictly exceeds the separation capacity $C_{\mathcal{H}}$, the Margin-Capacity Profile evaluates exactly to $1 - C_{\mathcal{H}}/\Gamma$, which strictly bounds the penalty between 0 and 1. While not diverging to infinity, this saturation causes the approximation gap to lower-bound away from zero, rendering the 0-1 error consistency guarantee effectively vacuous.

By making the margin $\Gamma(s_t, a_t)$ functionally dependent on the estimated local semantic density (Equation 14), SAMPPO dynamically optimizes this Margin-Capacity Profile. It actively minimizes the extent to which Γ exceeds the collapsing local capacity $C_{\mathcal{H}} \to 0$ in dense synonym clusters, constraining the policy against unaligned variance. SAMPPO minimizes, rather than strictly eliminates, the approximation gap in these extreme edge cases, keeping the structural requirements for $\mathcal{H}$-consistency as tight as possible.

G Implementation Details of SAMPPO and Baselines

To accommodate space constraints in the main body, this section provides comprehensive implementation details for the baselines, data configurations, and the variance-reduction techniques required to translate the continuous theoretical formulation of SAMPPO into a discrete, batch-wise optimization loop.

Models, Datasets, and Reward Model. We evaluate SAMPPO using two distinct foundation models: `Meta-Llama-3.1-8B-Instruct` [Llama Team, AI @ Meta, 2024] and `gemma-2-2b-it` [Gemma Team, Google DeepMind, 2024]. For the reward signal, we use a state-of-the-art dense model using the DeBERTaV3 architecture [He et al., 2023], `OpenAssistant/reward-model-deberta-v3-large-v2`, which is trained for conversational preference ranking [Köpf et al., 2023]. The models are trained and evaluated on subsets of the Anthropic HH-RLHF dataset [Bai et al., 2022], and the NVIDIA HelpSteer dataset [Wang et al., 2024b]. We extract 2,000 paired training queries from each dataset. We hold out random test subsets of $N = 500$ prompts for the final evaluations. All evaluations report the mean and standard deviation across 5 random *generation seeds* (fixed trained model, varied inference stochasticity). Pairwise statistical significance ($p < 0.05$) was assessed via a two-sided paired t-test on per-seed win rates.

Training Configuration. To enable efficient fine-tuning, the policy models are adapted using 4-bit Quantized Low-Rank Adaptation (QLoRA) [Dettmers et al., 2023] with the NF4 data type, double quantization, and `bfloat16` compute precision. The LoRA [Hu et al., 2022] adapters are applied to the query and value projection modules (`q_proj`, `v_proj`) with a rank of $r = 16$, $\alpha = 32$, and a dropout rate of 0.05. All methods use standard hyperparameters commonly adopted in the `trl` library [von Werra et al., 2020]. Specifically, we used a learning rate of 1.41×10^{-5} and an effective batch size of 32 (mini-batch size of 2 and 16 gradient accumulation steps). Generative rollouts during training use a maximum of 64 new tokens with standard sampling (nucleus sampling disabled via top_p = 1.0, temperature = 1.0). Standard PPO uses a fixed clipping parameter of $\epsilon = 0.20$. To ensure an apples-to-apples comparison, the BAPO empirical baseline [Xi et al., 2025] uses the exact same dynamic clipping bounds ($\epsilon_t \in [0.10, 0.30]$). Neither method uses an explicit entropy bonus ($c_2 = 0.0$). The KL divergence reward penalty anchoring the policy to the SFT checkpoint is set to $\beta_{\text{KL}} = 0.02$, and Generalized Advantage Estimation uses discount $\gamma = 0.99$ and decay $\lambda = 0.95$ (all other optimizer settings follow the `trl` library defaults).

Optimization was capped at a maximum of 100 optimization steps (effectively evaluating 3,200 trajectories). Since the training dataset subsets contain exactly 2,000 queries, 100 steps correspond to ≈ 1.6 epochs. In LLM RLHF literature, training beyond 1–2 epochs is widely known to induce severe over-optimization (the "alignment tax") and mode collapse. Stopping at 100 steps captures the critical early learning window before standard PPO over-optimizes the reward model and suffers generative collapse (as shown in Figure 2).

Offline Baselines Setup. For HH-RLHF, the DPO and SimPO baselines were trained on 2,000 human-labeled preference pairs provided directly by the dataset (i.e., these offline methods were optimized on ground-truth human labels, whereas the online methods were optimized against the DeBERTaV3 proxy reward model; this difference in reward signal quality is an inherent methodological asymmetry between the two paradigms and may partially favor the offline baselines). For HelpSteer, we generated candidate responses from the same base policy used in the online RLHF loop. Binary preference labels (chosen/rejected) were assigned by the DeBERTaV3 reward model, ensuring reward signal parity with the online methods. In both cases, the offline baselines received exactly 2,000 training pairs, matching the online methods' data budget. DPO was trained for 3 epochs with a learning rate of 1×10^{-5} and a uniform KL penalty $\beta = 0.1$. SimPO was trained for 3 epochs with a learning rate of 1×10^{-6}, reward scaling $\beta = 2.5$, and a relative target margin ratio of $\gamma/\beta = 0.55$.

Symmetrically Anchored Bounds: To prevent "gradient starvation" and ensure SAMPPO shares a comparable expected regularization intensity to the baseline, we symmetrically anchor our adaptive clipping bounds around standard PPO's fixed parameter ($\epsilon = 0.20$), dynamically modulating $\epsilon_t \in [0.10, 0.30]$.

Deterministic Pre-Computation (Variance Reduction): Standard PPO optimizes the surrogate objective over multiple epochs per batch. Relying on stochastic Monte Carlo action sampling to compute $D_{\pi_{\text{old}}}(s_t, a_t)$ during these epochs introduces variance, causing the clipping bound ϵ_t to fluctuate randomly on every backward pass for the exact same state. To enforce a stationary surrogate

objective and eliminate gradient routing leakage, SAMPPO calculates the expectation *once* outside the epoch loop during the trajectory rollout, using the deterministic top-k action embeddings evaluated from the fixed behavior policy.

Choice of k = 5 (The Linguistic Branching Factor): We select $k = 5$ because it aligns with the probabilistic distribution of auto-regressive language generation. In typical contexts, a confident language model's probability mass is concentrated in a 'nucleus' of roughly 3 to 5 valid semantic continuations (e.g., synonyms or syntactically valid alternative words). Sampling $k = 5$ evaluates the conditional semantic isolation within this plausible linguistic head while avoiding the computational overhead of evaluating the full vocabulary. Setting k too low ($k = 3$) suffers from high sampling variance and fails to capture valid synonyms, while setting k too high ($k = 10$) dilutes the isolation score by comparing the target action against noisy, improbable tokens.

Metric-Bounded Absolute Normalization: In typical RLHF pipelines, hardware constraints limit mini-batch sizes (e.g., to 2). Computing dynamic min-max normalization over such small mini-batches introduces extreme statistical variance that destabilizes the trust region. To enforce a stationary surrogate objective and ensure the scale is consistently determined across all samples for all time steps, we substitute batch normalization with *Absolute Normalization*.

Choice of $D_{\text{bound}} = 0.3$ (The Anisotropy Bound): We select $D_{\text{bound}} = 0.3$ to reflect the geometric properties of the LLM's representation space. It is well-documented that LLM embeddings are highly anisotropic, occupying a narrow cone in the vector space [Ethayarajh, 2019]. Empirically, the cosine distance ($1 - \cos$) between any two plausible, grammatically valid tokens in our semantic space is tightly bounded near 0.3 (i.e., cosine similarities range between 0.7 and 1.0). By setting $D_{\text{bound}} = 0.3$, we map the upper bound of the semantic embedding cone to the maximum trust region relaxation, ensuring an absolute metric normalization that respects the model's geometry.

Computational Efficiency. A common concern is that calculating expected semantic distances at every optimization step introduces computational bottlenecks. Since we calculate ϵ_t statically once prior to the epoch loop, deterministically sample a narrow neighborhood ($k = 5$), and reuse the token embedding matrix already materialized in VRAM during the forward pass (the input embeddings in our base configuration, or the `lm_head` unembedding vectors in the output-embedding variant), the overhead reduces to a parallelizable 1×5 vector of distances per sampled token. Empirically, SAMPPO introduces a negligible $< 2\%$ **overhead** in wall-clock training time per batch compared to standard PPO.

Value Function Clipping. The adaptive ϵ_t is applied exclusively to the policy surrogate objective. The value function loss uses the standard fixed clipping parameter ($\epsilon_{\text{vf}} = 0.20$), as is the default in the `trl` library.

Compute Resources. All model training and empirical evaluations were conducted using hardware accelerators, specifically compute nodes equipped with NVIDIA A100 Tensor Core GPUs. The total computational budget for the experiments, including hyperparameter search, baseline evaluations, and the final multi-seed runs, required approximately 400 GPU hours.

G.1 Baseline Tuning Protocol

The comparison against BAPO, DPO, and SimPO is central to our empirical claim, so we enforced strict parity using identical dataset splits, reward models, and compute budgets, and verified that no baseline was under-tuned. Regarding the compute budget, the wall-clock profiling of Appendix K shows SAMPPO introduces a practically negligible $+1.9\%$ overhead in total step time compared to standard PPO, so empirical improvements are strictly due to geometric robustness rather than an inflated compute budget. To ensure the offline and adaptive baselines were not under-tuned, we rigorously replicated the original authors' hyperparameter search grids on a 10% held-out validation set:

DPO. We swept the KL penalty $\beta \in \{0.05, 0.1, 0.2\}$ and learning rates from 5×10^{-6} to 2×10^{-5}, selecting $\beta = 0.1$ and LR $= 1 \times 10^{-5}$ based on validation reward-vs-KL Pareto frontiers.

SimPO. We used the authors' recommended length-normalized formulation ($\beta = 2.5$) and swept the target margin ratio $\gamma/\beta \in \{0.5, 0.55, 0.6\}$, selecting the optimal 0.55 to maximize sequence generation quality.

BAPO. To strictly isolate our *semantic* metric's validity against their *empirical* entropy heuristic, we enforced the exact same adaptive clipping range ($\epsilon_t \in [0.10, 0.30]$), mini-batch size, and GAE parameters as SAMPPO. While matching the clipping range ensures a strict 1:1 ablation of the adaptive metrics, we additionally verified on our validation set that BAPO's internal entropy-balancing threshold heuristics were optimally tuned, ensuring the baseline was not artificially bottlenecked.

H Ablation Study

Table 3: Ablation study on hyperparameter robustness and embedding sensitivity. Win rates are calculated head-to-head against the Standard PPO baseline ($\epsilon = 0.20$) and evaluated by GPT-4o on the Anthropic HH-RLHF and NVIDIA HelpSteer test sets. Unlike Table 2, which averages over 5 random generation seeds, each ablation row is evaluated on a single generation seed; this accounts for the minor difference in the shared base configuration (57.2% here vs. $57.7\% \pm 2.6$ in Table 2, well within one standard deviation).

	Anthropic HH-RLHF			NVIDIA HelpSteer		
SAMPPO Configuration	**SAMPPO Win**	**PPO Win**	**Tie Rate**	**SAMPPO Win**	**PPO Win**	**Tie Rate**
Hyperparameter: Top-k candidates (Fixed D_{bound} = 0.3)						
$k = 3$	56.6%	41.8%	1.6%	44.0%	47.4%	8.6%
$k = 5$ (Optimal)	57.2%	40.6%	2.2%	45.8%	41.2%	13.0%
$k = 10$	51.6%	46.4%	2.0%	45.0%	44.8%	10.2%
Hyperparameter: Absolute Metric Bound D_{bound} (Fixed $k = 5$)						
$D_{\text{bound}} = 0.2$	49.2%	47.6%	3.2%	42.0%	46.8%	11.2%
$D_{\text{bound}} = 0.3$ (Optimal)	57.2%	40.6%	2.2%	45.8%	41.2%	13.0%
$D_{\text{bound}} = 0.4$	46.4%	51.8%	1.8%	46.6%	44.0%	9.4%
Embedding Representation Space ($k = 5$, D_{bound} = 0.3)						
Input Embeddings (Table 2 Base)	57.2%	40.6%	2.2%	45.8%	41.2%	13.0%
Output Embeddings (`lm_head`)	**60.4%**	38.4%	1.2%	**46.6%**	41.6%	11.8%
Discrete Metric (Fallback)	47.8%	50.2%	2.0%	44.8%	46.8%	8.4%

I Validation Study of the Semantic Isolation Score

This section details the validation study summarized in Section 5.2, establishing that the Semantic Isolation Score captures contextual semantic density rather than policy confidence, token frequency, or embedding anisotropy.

I.1 Qualitative Examples of Decoding States

We systematically extracted and analyzed decoding states across the HH-RLHF test set using Llama-3.1-8B in `bfloat16`. Because we apply an absolute metric normalization ($D_{\text{bound}} = 0.3$), the score explicitly controls for raw embedding scale. Representative examples from both extremes:

High-density states (SAMPPO correctly tightens the clip to prevent over-optimizing noise among indistinguishable synonyms):

- (Score: 0.0812) Context: "The old building was completely abandoned, empty, and" — Top-5: [' vacant', ' deserted', ' unoccupied', ' desolate', ' bare'].
- (Score: 0.0935) Context: "The student was incredibly smart, highly intelligent, and exceptionally" — Top-5: [' brilliant', ' clever', ' bright', ' sharp', ' gifted'].
- (Score: 0.1140) Context: "The comedian's performance was absolutely hilarious and incredibly" — Top-5: [' funny', ' amusing', ' comical', ' entertaining', ' hysterical'].

Low-density states (SAMPPO correctly identifies the sparse semantic space and relaxes the clip for confident optimization):

- (Score: 0.6104) Context: "When you encounter this critical server error, you should immediately" — Top-5: [' restart', ' contact', ' check', ' stop', ' report'] (distinct operational actions).

- (Score: 0.6421) Context: "To improve your cardiovascular health, the doctor recommended that you" — Top-5: [' exercise', ' eat', ' stop', ' sleep', ' take'] (distinct lifestyle concepts).
- (Score: 0.5982) Context: "The primary cause of the economic recession was the sudden drop in" — Top-5: [' housing', ' consumer', ' global', ' interest', ' oil'] (distinct macroeconomic factors).

Notice that the policy exhibits similar uncertainty (spreading probability mass) in both sets of examples, so the metric does not merely capture raw policy confidence. Furthermore, the tokens within these synonym clusters span both high and low frequencies in the pre-training corpus, indicating that the metric isolates contextual semantic density rather than uncalibrated token frequency.

I.2 Corpus-Level Systematic Analysis

Because hand-selected examples cannot establish systematicity, we ran a corpus-level study over all 28,413 decoding states from the 500 held-out generations used in our expanded evaluation (Llama-3.1-8B, HH-RLHF):

Not confidence. The Spearman correlation between the isolation score and the policy's entropy is $\rho = 0.21$ (top-1 probability: $\rho = -0.17$): a weak association, confirming the score is largely independent of policy confidence.

Not frequency. The correlation with the mean pre-training log-frequency of the top-5 candidates is $\rho = -0.12$.

External synonym labels (embedding-independent ground truth). We label each state by whether at least 3 of the top-5 candidates share a WordNet [Miller, 1995] synset or direct near-synonym set (tokens normalized by stripping leading whitespace and lowercasing before lemma matching). The isolation score separates synonym-cluster states (11.8% of the corpus) from distinct-candidate states with AUC = 0.94.

Anisotropy control. Absolute normalization calibrates the scale of the score but does not eliminate anisotropy. We therefore recomputed all scores after the standard correction of Mu and Viswanath [2018], mean-centering and removing top principal components (stable for removing 1–40 components, evaluated at $\{1, 2, 3, 5, 10, 20, 40\}$; we report 3). Corrected and original scores correlate at $\rho = 0.93$ and the separation is preserved (AUC = 0.92): the signal is not an anisotropy artifact.

J Extended Objective Evaluation

This section details the evaluation protocols summarized in Section 5, which complement the primary GPT-4o judge with objective, judge-free metrics and a second independent judge. All evaluations use the same N=500 held-out generations (Llama-3.1-8B, HH-RLHF); results are collected in Table 4.

Table 4: Objective evaluation beyond a single judge (Llama-3.1-8B, HH-RLHF, N=500): ArmoRM reward model [Wang et al., 2024a], Distinct-N diversity, and head-to-head rates under the second judge GPT-4-Turbo (86.4% agreement with GPT-4o).

				GPT-4-Turbo H2H vs. SAMPPO		
Algorithm	**ArmoRM (↑)**	**Distinct-2 (↑)**	**Distinct-3 (↑)**	**SAMPPO Win**	**Baseline Win**	**Tie**
PPO Baseline	0.0415 ± 0.0018	0.8124 ± 0.0231	0.8540 ± 0.0184	**55.8%**	41.2%	3.0%
BAPO Baseline	0.0431 ± 0.0012	0.8915 ± 0.0167	0.9320 ± 0.0113	**53.6%**	39.4%	7.0%
SAMPPO	**0.0449 ± 0.0009**	**0.9492 ± 0.0082**	**0.9785 ± 0.0041**	–	–	–

Objective reward-model evaluation (ArmoRM). We scored every response with the state-of-the-art open-source reward model `RLHFlow/ArmoRM-Llama3-8B-v0.1` [Wang et al., 2024a], which is interpretable, multi-objective, and, importantly, independent of the DeBERTaV3 reward model used during training, guarding against reward over-optimization artifacts. SAMPPO achieved a superior average score of 0.0449, strictly outperforming both standard PPO (0.0415) and the BAPO empirical baseline (0.0431).

Sequence-level diversity (Distinct-N). To objectively measure the mitigation of entropy collapse at the sequence level, rather than only via token-level policy entropy, we computed Distinct-2 and Distinct-3 [Li et al., 2016] (the proportion of unique bi-/tri-grams). SAMPPO preserves stable sequence-level diversity (Distinct-2: 0.9492, Distinct-3: 0.9785) versus BAPO (0.8915, 0.9320) and standard PPO, which collapsed into repetitive, low-entropy generation (0.8124, 0.8540).

Second independent judge (GPT-4-Turbo). Re-evaluating the full N=500 test set with GPT-4-Turbo (temperature 0, randomized presentation order), SAMPPO maintained strong, statistically significant win rates against both PPO (55.8% Win, 41.2% Loss, 3.0% Tie) and BAPO (53.6% Win, 39.4% Loss, 7.0% Tie), with an 86.4% overall per-comparison agreement rate with GPT-4o. A two-sided sign test on the per-example preferences (ties excluded) rejects the null hypothesis of no preference in both cases: 279 wins vs. 206 losses against PPO (n = 485) and 268 vs. 197 against BAPO (n = 465), each with $p \approx 0.001$.

K Wall-Clock Profiling

To conclusively quantify the cost-benefit tradeoff against adaptive clipping heuristics such as BAPO, we measured real wall-clock execution times for a full optimization step, auto-regressive generation, GAE computation, and 4 PPO epochs of surrogate backward passes, on an NVIDIA A100 GPU (batch size 32, sequence length 64, vocabulary size 128,256).

Because the behavior policy's top-k probabilities and the token embedding matrix (input or `lm_head`) are already computed and resident in VRAM during the PPO forward pass, SAMPPO requires only a simple $1 \times k$ cosine similarity calculation per generated token. The isolated surrogate loss computation per mini-batch is extremely fast: 0.0076 s for standard PPO vs. 0.0102 s for SAMPPO. Accumulated across the entire training step, which averages 8.50 s in standard PPO due to auto-regressive generation and weight updates, SAMPPO's total step time increases to just 8.66 s. This yields a measured, end-to-end step-time overhead of only +1.9%, with zero measurable extra memory footprint. Unlike BAPO-style heuristics, this modest constant is accompanied by the formal guarantees of Section 4; the stability gains are computationally almost free.

L SAMPPO vs. Global Entropy Regularization

A natural question arises from the observation of PPO's entropy collapse: why not simply increase the standard PPO entropy bonus coefficient ($c_2 \cdot H(\pi)$) to force exploration?

While a naive entropy bonus algebraically increases $H(\pi)$, it is fundamentally *structure-agnostic*. A global entropy bonus forces the policy to allocate probability mass to all actions uniformly, unconditionally amplifying irrelevant, structurally incoherent, or toxic tokens alongside valid synonyms. This degrades alignment and context relevance to force spurious diversity.

In contrast, SAMPPO acts as a *local, geometry-aware regularizer*. By dynamically scaling the trust region based on local semantic density, it prevents probability mass from collapsing onto a single synonym, preserving semantic diversity among valid linguistic candidates *without* forcing the model to hallucinate or allocate mass to irrelevant outliers.

M Broader Societal Impacts

Our work introduces SAMPPO, which improves the alignment of Large Language Models (LLMs) via reinforcement learning from human feedback (RLHF). On the positive side, enhancing the stability and generative diversity of aligned models makes them safer, more reliable, and better equipped to faithfully follow human intent without suffering from mode collapse. However, as with any advancement in alignment and fine-tuning efficiency, there is a potential dual-use risk. Malicious actors could theoretically leverage such optimized, sample-efficient frameworks to fine-tune models toward harmful objectives, such as generating persuasive disinformation or toxic content. To mitigate these risks, our research and evaluations rely strictly on established, safety-focused alignment benchmarks (e.g., Anthropic HH-RLHF). We strongly advocate for the continued development of robust safety guardrails and careful curation of preference data when deploying these systems.