Trust Region Policy Distillation

Zhengpeng Xie[1*] **Li Lyna Zhang**[3,†] **Zeke Xie**[1,2,‡] **Mao Yang**[3]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]Shanghai Xingyunzhili Artificial Intelligence Institute
zhengpengxie00@gmail.com li.zhang@xingyunzhili.com
zekexie@hkust-gz.edu.cn maoyang@xingyunzhili.com

Our code is available at: https://github.com/xie-lab-ml/TOP-D

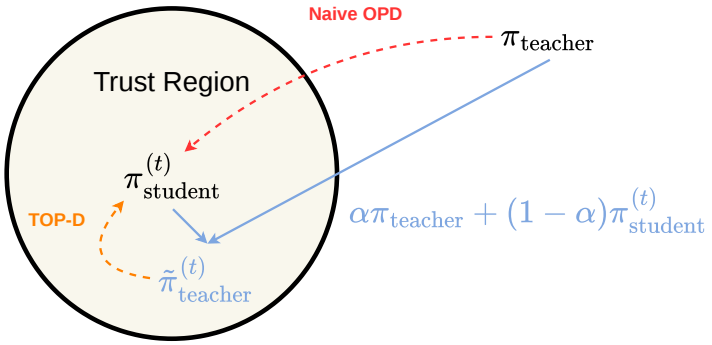

Figure 1: While standard OPD forces direct teacher supervision and suffers from unstable optimization, TOP-D ensures stable training by dynamically constructing a proximal teacher at every iteration.

Abstract

Big goals are hard to achieve all at once; breaking them into small steps is wiser. We present Trust Region Policy Distillation (TOP-D), which transforms the notoriously unstable, high-variance On-Policy Distillation (OPD) into a stable training paradigm by dynamically constructing a *proximal teacher*. Theoretically, we establish a rigorous framework demonstrating that TOP-D inherently controls gradient variance. By providing a formal global convergence analysis alongside a monotonic improvement bound, we mathematically formalize the reliability and stability of the overall training dynamics. Empirically, TOP-D dramatically enhances training stability, sample efficiency, and final performance on mathematical reasoning tasks. More importantly, TOP-D introduces zero additional computational overhead, positioning itself as a promising alternative to the well-established OPD paradigm.

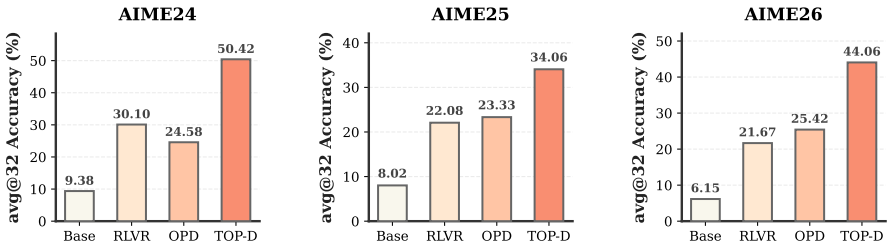

Figure 2: The avg@32 accuracy on several AIME benchmarks for RLVR applied to the Qwen3-8B-Base model, and for OPD and TOP-D using Qwen3-30B-A3B-Instruct-2507 as the teacher model.

*Did this work during the internship at Xingyunzhili. †Project leader. ‡Corresponding author.

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

Method	Sampling	Reward Signal	Training Stability	Theoretical Guarantee
Supervised Fine-Tuning	off-policy	**dense**	**high**	weak
Reinforcement Learning with Verifiable Rewards	**on-policy**	sparse	low	weak
On-Policy Distillation	**on-policy**	**dense**	low	weak
Trust Region Policy Distillation	**on-policy**	**dense**	**high**	**strong**

Table 1: Comparison of different paradigms.

1 Introduction

On-Policy Distillation (OPD) [12] has rapidly emerged as the prevailing paradigm for post-training Large Language Models (LLMs) [22, 19, 25, 23, 2], in which the student learns from token-level signals provided by an external teacher. Mathematically, this is equivalent to reinforcement learning where the immediate reward is defined by the logarithmic probability ratio between the teacher and student policies. Consequently, the on-policy nature of OPD prevents the catastrophic forgetting of Supervised Fine-Tuning (SFT) [17], while its dense rewards overcome the sample inefficiency and instability of Reinforcement Learning with Verifiable Rewards (RLVR) (as summarized in Table 1).

Despite its theoretical appeal, standard OPD suffers from severe optimization fragility in practice. The fundamental bottleneck of OPD lies in the inherent capacity gap between the target teacher and the student [7]. Because the standard distillation objective is driven by the logarithmic probability ratio between the two policies, the optimization dynamics are highly vulnerable to teacher-student disagreements. Recent studies have devoted considerable engineering efforts to mitigate this instability through various techniques, including teacher-student mixed sampling [5], reward clipping [9], top-p sampling [3], off-policy cold starts [11], and scaling to full-vocabulary supervision [2]. Yet these mitigation strategies remain largely empirical heuristics that fail to capture the underlying mechanism of the instability: the unbounded logarithmic probability difference between the teacher and student. Furthermore, these empirical methods inherently lack theoretical guarantees for reliable optimization.

In this paper, we present Trust Region Policy Distillation (TOP-D) to systematically resolve the inherent optimization instability of standard OPD. Specifically, to overcome the unbounded logarithmic signals, we dynamically construct an *external proximal teacher* (as illustrated in Figure 1 and Section 3.1) by interpolating the probability distributions of the target teacher and the current student. This mathematical construction transforms the unbounded distillation reward into a smooth, strictly lower-bounded signal, analytically preventing variance explosion. Building upon this stabilized objective, we further incorporate *internal trust region iterations* (described in Section 3.2) to better approximate the current proximal teacher, while simultaneously improving overall sample efficiency.

Crucially, while the proximal teacher is constructed conceptually, the practical objective mathematically reduces to a simple *algebraic transformation* of the standard token-level OPD reward. As a result, TOP-D introduces zero additional computational overhead, serving as a plug-and-play module that is natively compatible with both state-of-the-art distributed training infrastructures [18, 27] and high-throughput inference engines [10, 28]. Beyond this empirical elegance, our proposed framework establishes a rigorous theoretical closed-loop: while the external proximal teacher strictly bounds the gradient variance (described in Section 4.1), the internal trust region iterations guarantee monotonic policy improvement (described in Section 4.3), systematically minimizing the single-step optimization error to tightly close the overall asymptotic convergence gap (described in Section 4.2).

In summary, the main contributions of this paper are as follows:

- We propose TOP-D, a plug-and-play distillation paradigm that smooths the unbounded reward signal and safely breaks the strict on-policy data-reuse barrier of standard OPD.
- We establish a rigorous theoretical framework where the proximal teacher inherently bounds gradient variance to stabilize training, while our convergence analysis and monotonic improvement bound form a closed-loop to systematically minimize the asymptotic error.
- We empirically demonstrate that TOP-D dramatically improves reasoning capabilities on mathematical reasoning tasks, delivering a 25.84% absolute improvement in avg@32 accuracy over standard OPD on AIME24 and consistently dominating RLVR and OPD baselines.

2 Preliminaries

2.1 On-Policy Distillation

Suppose we have an autoregressive language model π_θ, where $\theta \in \Theta$ denotes the parameters. Given a prompt $x \sim \mathcal{D}_x$, the model generates a response $y = (y_0, y_1, \ldots, y_T)$[2] autoregressively, where $T = |y|$ is the response length. Formally,

$$\log \pi_\theta(y \mid x) = \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}), \tag{1}$$

where $y_{<t}$ denotes $(y_0, y_1, \ldots, y_{t-1})$. In the context of OPD, we aim to minimize the following *reverse* KL divergence:

$$\begin{aligned} \theta^* = \arg\min_\theta \mathcal{L}(\theta) &= \arg\min_\theta \mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}_{\mathrm{KL}}(\pi_\theta(\cdot \mid x), \pi^*(\cdot \mid x))\right] \\ &= \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(\cdot|x)}\left[\log \frac{\pi^*(y \mid x)}{\pi_\theta(y \mid x)}\right] = \arg\max_\theta J(\theta), \end{aligned} \tag{2}$$

where π^* is typically a stronger model. We are now ready to present the standard *policy gradient* form of the objective (2):

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(\cdot|x)}\left[\sum_{t=1}^{|y|} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t}) \cdot \left(\sum_{k=t}^{|y|} \log \rho_k\right)\right], \tag{3}$$

where $\rho_k = \frac{\pi^*(y_k|x,y_{<k})}{\pi_\theta(y_k|x,y_{<k})}$. In other words, we can treat $\log \rho_k$ as the immediate reward for the k-th token, i.e., $r_k = \log \rho_k$. However, the cumulative reward $\sum_{k=t}^{|y|} r_k$ introduces an inherent length bias as well as high variance. In practice, a common approach is to use the immediate reward r_k directly.

2.2 Reinforcement Learning Formulation

To establish a rigorous theoretical foundation, we cast autoregressive language modeling as a deterministic Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mu_0)$.

- **State space $\mathcal{S}$ and action space $\mathcal{A}$.** State $s_t = (x, y_{<t})$ represents the prompt and prefix. The action space $\mathcal{A} = \mathcal{V}$ includes the vocabulary and the `EOS` token. To rigorously handle variable-length sequences, we augment $\mathcal{S}$ with a permanent *absorbing state* $s_\perp$.
- **Dynamics $\mathcal{P}$ and initial distribution μ_0.** The initial state distribution $\mu_0(s)$ follows the prompt distribution $\mathcal{D}_x$ and the special initial token y_0. Transitions are *deterministic*: selecting $a \neq$ `EOS` appends a to the prefix, while selecting `EOS` definitively transitions to $s_\perp$.
- **Reward function r.** The immediate reward is the distillation signal: $r(s, a) = \log \frac{\pi^*(a|s)}{\pi(a|s)}$. To prevent infinite accumulation after termination[3], we set $r(s_\perp, a) \equiv 0$ for any $a \in \mathcal{A}$.

Let $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ be a stochastic policy. In the undiscounted setting ($\gamma = 1$), the action value function $Q^\pi(s, a)$ is defined as $Q^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} r(s_{t+k}, a_{t+k}) \mid s_t = s, a_t = a\right]$. The state value function $V^\pi(s)$ is strictly obtained by marginalizing Q^π over the policy: $V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s, a)\right]$. Consequently, we define the advantage function, which forms the basis for policy improvement bounds, as $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$. To decouple sequential dependencies, we introduce the *unnormalized* state distribution $d^\pi(s) = \sum_{t=1}^{\infty} \mathbb{P}(s_t = s \mid \pi)$. The true objective of a policy π is the expected cumulative reward over the infinite horizon:

$$\eta(\pi) = \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi(\cdot|x)}\left[\sum_{t=1}^{\infty} r(s_t, a_t)\right]. \tag{4}$$

While this abstract MDP formalization is adopted for the theoretical proofs in Section 4.3, the remainder of the paper defaults to the notation established in Section 2.1.

[2] Here, y_0 denotes the special initial token and is not counted toward the response length.

[3] Assume that $\pi(\texttt{EOS} \mid s_\perp) = 1$.

3 Methodology

This section details the design of our core algorithm, Trust Region Policy Distillation (TOP-D). We begin in Section 3.1 by constructing a *proximal teacher* to stabilize training. Subsequently, in Section 3.2, the current student approximates this proximal teacher by performing *trust region iterations*. Finally, Section 3.3 outlines the practical implementation of the TOP-D algorithm.

3.1 External Proximal Teacher

The fundamental challenge in standard On-Policy Distillation (OPD) arises from the inherent capacity gap between the student policy π_θ and the target teacher policy π^*. As defined in Equation (3), the optimization dynamics heavily rely on the probability ratio $\rho_k = \frac{\pi^*(y_k|x,y_{<k})}{\pi_\theta(y_k|x,y_{<k})}$. In extreme cases where the teacher assigns a near-zero probability to a generated token ($\pi^* \to 0$), the immediate reward $r_k = \log \rho_k$ diverges to negative infinity. This unbounded penalty destabilizes the policy gradient, making the optimization process extremely fragile.

To address this instability, we draw inspiration from *trust region* methods in reinforcement learning [14, 15, 20], which restrict policy updates from taking destructively large steps. Rather than forcing the student to perfectly imitate a rigid teacher, we dynamically construct an intermediate, localized target distribution. We term this surrogate target the *proximal teacher* $\tilde{\pi}^*$. Formally,

$$\tilde{\pi}^*(y_k \mid x, y_{<k}) = \alpha\pi^*(y_k \mid x, y_{<k}) + (1-\alpha)\pi_\theta(y_k \mid x, y_{<k}),$$

where $\alpha \in (0, 1)$ is the coefficient that controls the interpolation strength. In fact, the proximal teacher inherently smooths the token-level reward r_k. Specifically,

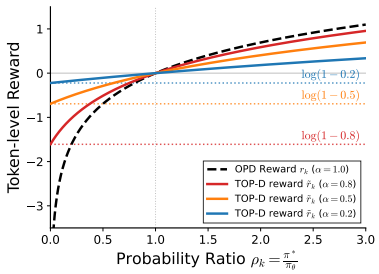

Figure 3: Reward curves of r_k and $\tilde{r}_k$.

$$\tilde{r}_k = \log \frac{\tilde{\pi}^*(y_k \mid x, y_{<k})}{\pi_\theta(y_k \mid x, y_{<k})} = \log\left(\alpha \frac{\pi^*(y_k \mid x, y_{<k})}{\pi_\theta(y_k \mid x, y_{<k})} + 1 - \alpha\right) = \log\left(\alpha\rho_k + 1 - \alpha\right). \quad (5)$$

When $\alpha = 1$, it reduces to the original OPD reward r_k. Thanks to mathematics, we do not need to explicitly construct a proximal teacher to obtain the new reward $\tilde{r}_k$. Figure 3 visualizes the OPD reward r_k and our proposed reward $\tilde{r}_k$ with respect to the probability ratio ρ_k. We can see that $\tilde{r}_k$ is strictly bounded from below[4] as $\rho_k \to 0$. Moreover, we explicitly justify our choice of interpolating in the *probability space* instead of the *log-probability space* to construct the proximal teacher, since setting $\log \tilde{\pi}^*(y_k \mid x, y_{<k}) = \alpha \log \pi^*(y_k \mid x, y_{<k}) + (1-\alpha) \log \pi_\theta(y_k \mid x, y_{<k})$ yields

$$\tilde{r}_k = \alpha r_k, \quad (6)$$

which is merely a naive scaling of the original OPD reward.

3.2 Internal Trust Region Iterations

With the proximal teacher established, the student policy is optimized to maximize the expected cumulative reward $\tilde{R}_t = \sum_{k=t}^{|y|} \tilde{r}_k$. While standard OPD relies on continuous online sampling, it fundamentally operates in a strictly on-policy manner. In the naive formulation, once the student policy is updated, the previously generated trajectories are immediately discarded. This results in notoriously low sample efficiency, as OPD lacks the capacity to reuse data.

Inspired by modern trust region algorithms [14, 15, 20], we analogously decouple the *behavior policy* $\pi_{\theta_{\text{old}}}$ and the *target policy* π_θ to inherit their high sample efficiency. Specifically,

$$\mathcal{J}(\theta) = \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_x \\ \{y^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|x)}} \left[\frac{1}{\sum_{i=1}^G |y^i|} \sum_{i=1}^G \sum_{t=1}^{|y^i|} \left\{ \min\left[p_t^i \cdot \hat{A}_t^i, \text{clip}\left(p_t^i, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t^i \right] \right\} \right]. \quad (7)$$

Here, we follow the group-based reinforcement learning settings [16, 6], where the superscript i indexes different model responses and $p_t^i = \frac{\pi_\theta(y_t^i|x,y_{<t}^i)}{\pi_{\theta_{\text{old}}}(y_t^i|x,y_{<t}^i)}$ is the importance sampling ratio. We

[4] $\log(\alpha x + 1 - \alpha) \geq \log(1-\alpha)$ when $x \in (0, +\infty)$ and $\alpha \in (0, 1)$.

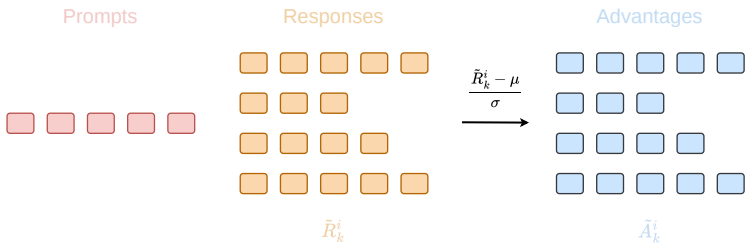

Figure 4: In contrast to *sequence-level* advantage normalization, we perform *token-level* normalization across the responses generated for a given prompt to better leverage the dense reward signal.

Algorithm 1 Trust Region Policy Distillation (TOP-D)

Require: Student π_θ, teacher π^*, dataset $\mathcal{D}_x$, interpolation α, group size G, internal epochs E

while not converged **do**
  Initialize $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Sample a batch of prompts $\mathcal{X} \subset \mathcal{D}_x$
  for each prompt $x \in \mathcal{X}$ **do**
    Perform rollout to sample G responses $\{y^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$
    Compute TOP-D rewards $\tilde{r}_k^i \leftarrow \log\left(\alpha \frac{\pi^*(y_k^i | x, y_{<k}^i)}{\pi_{\theta_{\text{old}}}(y_k^i | x, y_{<k}^i)} + 1 - \alpha\right)$
    Compute token-level returns $\tilde{R}_k^i \leftarrow \tilde{r}_k^i + \frac{1}{\max\{|y^i|-k,1\}} \sum_{j=k+1}^{|y^i|} \tilde{r}_j^i$
    Compute token-level advantages $\hat{A}_k^i \leftarrow \frac{\tilde{R}_k^i - \mu}{\sigma}$
  end for
  for epoch $e = 1$ to E **do**
    $\mathcal{J}(\theta) \leftarrow \frac{1}{\sum_{i=1}^G |y^i|} \sum_{i=1}^G \sum_{t=1}^{|y^i|} \left\{ \min\left[p_t^i \cdot \hat{A}_t^i, \text{clip}\left(p_t^i, 1-\epsilon, 1+\epsilon\right) \cdot \hat{A}_t^i \right] \right\}$
    Update θ by maximizing $\mathcal{J}(\theta)$ using an optimizer
  end for
end while

adopt a more fine-grained token-level normalized advantage $\hat{A}_t^i$ to better exploit dense reward signals, as illustrated in Figure 4. For each prompt, the advantage relies on a token-level return $\tilde{R}_k^i$, which combines the immediate reward and a length-normalized future return:

$$\tilde{R}_k^i = \tilde{r}_k^i + \frac{\sum_{j=k+1}^{|y^i|} \tilde{r}_j^i}{\max\{|y^i| - k, 1\}} = \begin{cases} \tilde{r}_k^i + \frac{1}{|y^i|-k} \sum_{j=k+1}^{|y^i|} \tilde{r}_j^i, & k < |y^i|, \\ \tilde{r}_k^i, & k = |y^i|. \end{cases} \tag{8}$$

Empty sums are zero. Empirically, this return discourages both overly short and excessively long responses. For any given prompt, the mean μ and standard deviation σ within each group are calculated via $\mu = \text{mean}\{\tilde{R}_k^i\}$ and $\sigma = \text{std}\{\tilde{R}_k^i\}$, where $i = 1, \ldots, G$ and $k = 1, \ldots, |y^i|$.

3.3 Trust Region Policy Distillation

In this section, we synthesize the concepts of the *external proximal teacher* in Section 3.1 and the *internal trust region iterations* in Section 3.2 to formulate our complete algorithmic framework, Trust Region Policy Distillation (TOP-D), summarized in Algorithm 1.

Standard OPD suffers from two fundamental bottlenecks: optimization fragility and low sample efficiency. TOP-D systematically resolves both through a unified design. First, the *external proximal teacher* transforms the unbounded OPD reward into a smooth, lower-bounded signal $\tilde{r}_k$, inherently stabilizing the optimization process. Building upon this principle, the *internal trust region iterations* decouple the behavior and target policies, breaking the strict on-policy data-reuse barrier. Crucially, these mechanisms are not merely empirical heuristics but share deep theoretical synergy. We rigorously formalize this framework and establish its mathematical guarantees in Section 4.

4 Theoretical Analysis

In this section, we establish a rigorous theoretical foundation for TOP-D. We begin in Section 4.1 by proving that the proximal teacher inherently bounds gradient variance to stabilize optimization. Building upon this, we derive a global convergence bound in Section 4.2 by abstracting the learning dynamics. Finally, to minimize the local error bottleneck identified in this convergence analysis, we establish the monotonic improvement guarantee for our internal trust region iterations in Section 4.3.

4.1 Bounded Variance

To rigorously analyze the gradient stability, we restrict our discussion to the autoregressive generation environment over a finite vocabulary $\mathcal{V}$. In policy gradient methods, the variance of the gradient is typically dominated by its *second moment*, i.e., $\mathrm{tr}(\mathrm{Var}(g)) \leq \mathbb{E}[\|g\|^2]$. Thus, our theoretical analysis will primarily focus on establishing an absolute upper bound for the gradient's second moment.

Assumption 4.1. We assume that the gradient of the student policy's log-probability (the score function) is uniformly bounded. That is, there exists a constant $M > 0$, such that for any prompt $x \sim \mathcal{D}_x$, generated response $y \sim \pi(\cdot \mid x)$, and k-th token y_k, we have

$$\|\nabla_\theta \log \pi_\theta(y_k \mid x, y_{<k})\| \leq M. \tag{9}$$

Let $g_k = \nabla_\theta \log \pi_\theta(y_k \mid x, y_{<k}) \cdot r_k$ denote the token-level policy gradient estimator for standard OPD, and $\tilde{g}_k = \nabla_\theta \log \pi_\theta(y_k \mid x, y_{<k}) \cdot \tilde{r}_k$ denote the gradient estimator for our proposed TOP-D.

Theorem 4.2. *Under Assumption 4.1, for any interpolation coefficient $\alpha \in (0, 1)$, the variance of the TOP-D gradient estimator $\tilde{g}_k$ does not diverge across the entire probability space. Instead, it is strictly bounded by a uniform absolute limit:*

$$\mathrm{tr}(\mathrm{Var}(\tilde{g}_k)) \leq M^2 \left(\log^2(1-\alpha) + C^* \alpha \right), \tag{10}$$

where C^* is a universal mathematical constant independent of the vocabulary size.

Proof. See Appendix E.1. □

Remark 4.3. TOP-D employs an asymmetric operator to stabilize training. For heavy penalizations, it introduces a hard safety bound, $\log^2(1-\alpha)$, preventing catastrophic divergence; for high rewards, it applies a linear dampening scaled by $C^*\alpha$. This effectively designates α as a direct variance controller. As $\alpha \to 1^-$, TOP-D recovers standard OPD, with its upper bound diverging ($\log^2(1-\alpha) \to \infty$) to reflect the lack of explicit variance control in OPD. Conversely, as $\alpha \to 0^+$, the variance vanishes.

4.2 Convergence Analysis

To rigorously analyze the convergence properties of TOP-D, we abstract the construction of the proximal teacher as an iterative operator mapping over the policy space. Let Π denote the space of all valid policies. To quantify the divergence within this space, we introduce the expected L_1 distance between any two policies π_i, π_j as $d(\pi_i, \pi_j) = \mathbb{E}_{x \sim \mathcal{D}_x} [\|\pi_i(\cdot \mid x) - \pi_j(\cdot \mid x)\|_1]$. Next, we formalize the iterative process through the *proximal teacher operator* $\mathcal{T} : \Pi \to \Pi$, given by

$$\mathcal{T}(\pi) = \alpha \pi^* + (1-\alpha)\pi, \tag{11}$$

where $\alpha \in (0, 1)$ is the interpolation coefficient, and π^* is the target teacher. Building upon this metric space, we are now ready to establish the following convergence analysis:

Theorem 4.4. *For any initial student policy $\pi_0 \in \Pi$, consider the iterative update process $\pi_{k+1} = \mathcal{T}(\pi_k) + \epsilon_k$[5], where ϵ_k accounts for the practical optimization error (e.g., function approximation and optimization noise) at the k-th global step, where $k \geq 0$. Then, the following bound holds:*

$$d(\pi_{k+1}, \pi^*) \leq (1-\alpha)^{k+1} d(\pi_0, \pi^*) + \sum_{i=0}^{k} (1-\alpha)^{k-i} \|\epsilon_i\|_1. \tag{12}$$

Furthermore, assuming the optimization process eventually stabilizes such that the asymptotic error is bounded by $\limsup_{k\to\infty} \|\epsilon_k\|_1 \leq \epsilon_\infty$, we have

$$\limsup_{k \to \infty} d(\pi_{k+1}, \pi^*) \leq \frac{\epsilon_\infty}{\alpha}. \tag{13}$$

[5] With a slight abuse of notation, let $\|\epsilon_k\|_1 = d(\pi_{k+1}, \tilde{\pi}_k^*)$, where $\tilde{\pi}_k^* = \mathcal{T}(\pi_k)$ is the proximal teacher at the k-th step.

Proof. See Appendix E.2. □

Remark 4.5. The bound presented in Theorem 4.4 not only reveals an inherent *error forgetting* mechanism that isolates early updates, but also highlights a fundamental *theoretical bottleneck*: the asymptotic convergence gap is bounded by ϵ_∞/α. This implies that to break through the precision ceiling without compromising stability, i.e., keeping α small (we offer more details in Appendix C.1), the only viable path is to strictly minimize the single-step optimization error $\|\epsilon_k\|_1$. This theoretical bottleneck motivates us to introduce *internal trust region iterations* in the subsequent section.

4.3 Monotonic Improvement

As revealed in Theorem 4.4, the fundamental theoretical bottleneck of the proximal teacher framework lies in the asymptotic convergence gap bounded by ϵ_∞/α. To tightly close this gap, it is imperative to rigorously minimize the single-step optimization error $\|\epsilon_k\|_1$. TOP-D achieves this via internal trust region iterations. In this section, we establish the theoretical insight of this mechanism by proving monotonic policy improvement. For brevity, we follow the same notation established in Section 2.2.

Assumption 4.6. Given the fixed context window constraints inherent to autoregressive language models, we assume that any valid policy $\pi \in \Pi$ generates the `EOS` token within a finite maximum horizon $T_{\max}$. Consequently, the response length strictly satisfies $|y| \le T_{\max}$, so $\sum_{s \neq s_\perp} d^\pi(s) < \infty$.

In our undiscounted setting ($\gamma = 1$), the unnormalized measure $d^\pi(s)$ diverges because the permanent absorbing state $s_\perp$ dominates as $t \to \infty$. To construct a valid probability distribution over the meaningful transient states, we explicitly exclude $s_\perp$ and normalize $d^\pi(s)$.

Definition 4.7. We formally define the *expected response length* of a policy π as $\ell_\pi = \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi(\cdot|x)}[|y|]$. Excluding $s_\perp$, the sum of the state distribution precisely evaluates to this length. Thus, we define the normalized state visitation measure $d^\pi_{\text{norm}}(s) = \frac{d^\pi(s)}{\ell_\pi}$ for $s \neq s_\perp$.

Remark 4.8. This normalized measure enables an equivalent reformulation of the *Performance Difference Lemma* [8], strictly decoupling the sequence length from the single-step generation quality.

We can thus rewrite the original objective (4) as $\eta(\pi) = \ell_\pi \cdot \mathbb{E}_{s\sim d^\pi_{\text{norm}}(\cdot), a\sim\pi(\cdot|s)}[r(s,a)]$, and we are now ready to present the main theoretical result of this section:

Theorem 4.9. *Given Assumption 4.6 and Definition 4.7, for any two policies, $\tilde\pi$ and π, the following bound holds:*

$$\eta(\tilde\pi) \ge \zeta_\pi(\tilde\pi) - 2\xi T_{\max}\ell_\pi \mathbb{E}_{s\sim d^\pi_{\text{norm}}(\cdot)}\left[\mathcal{D}_{\text{TV}}(\tilde\pi(\cdot \mid s), \pi(\cdot \mid s))\right], \tag{14}$$

where $\zeta_\pi(\tilde\pi) = \eta(\pi) + \ell_\pi \mathbb{E}_{s\sim d^\pi_{\text{norm}}(\cdot), a\sim\tilde\pi(\cdot|s)}[A^\pi(s,a)]$, and $\xi = \max_s |\mathbb{E}_{a\sim\tilde\pi(\cdot|s)}[A^\pi(s,a)]|$.

Proof. See Appendix E.3. □

Remark 4.10. Let $\mathcal{M}_\pi(\tilde\pi)$ denote the lower bound established in Theorem 4.9. By definition, $\eta(\pi) = \mathcal{M}_\pi(\pi)$. Therefore, any policy update that improves the lower bound guarantees monotonic improvement on the true objective, i.e., $\eta(\tilde\pi) - \eta(\pi) \ge \mathcal{M}_\pi(\tilde\pi) - \mathcal{M}_\pi(\pi)$.

As a result, optimizing $\mathcal{M}_\pi(\tilde\pi)$ ensures that the original reverse KL divergence objective monotonically improves during the internal updates. Let the superscript (n) denote the n-th step of the internal trust region iterations at the k-th global step, and let $\tilde\pi_k^* = \mathcal{T}(\pi_k)$ be the fixed proximal teacher. By iteratively optimizing $\mathcal{M}_\pi(\tilde\pi)$, we obtain a monotonically decreasing sequence of KL divergences:

$$\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}^{(0)}_{\text{KL}}\right] \ge \mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}^{(1)}_{\text{KL}}\right] \ge \cdots \ge \mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}^{(n)}_{\text{KL}}\right], \tag{15}$$

where $\mathcal{D}^{(j)}_{\text{KL}} \triangleq \mathcal{D}_{\text{KL}}(\pi_k^{(j)}(\cdot \mid x), \tilde\pi_k^*(\cdot \mid x))$ denotes the reverse KL divergence objective at the j-th internal iteration. If the internal trust region optimization proceeds until the final expected divergence satisfies $\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}^{(n)}_{\text{KL}}\right] \le \frac{1}{2}\delta^2$ for $\delta > 0$, then by combining Pinsker's inequality with Jensen's inequality: $\|\epsilon_k\|_1 = d(\pi_{k+1}, \tilde\pi_k^*) = d(\pi_k^{(n)}, \tilde\pi_k^*) \le \mathbb{E}_{x\sim\mathcal{D}_x}\left[\sqrt{2\mathcal{D}^{(n)}_{\text{KL}}}\right] \le \sqrt{2\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}^{(n)}_{\text{KL}}\right]} \le \delta$, we rigorously guarantee that the single-step optimization error satisfies $\|\epsilon_k\|_1 \le \delta$. This result perfectly closes the overall theoretical loop: the internal trust region iterations mathematically guarantee the continuous minimization of the single-step optimization error ϵ_k, thereby enabling our proposed algorithm to tightly close the asymptotic convergence gap ϵ_∞/α identified in Theorem 4.4.

5 Experiments

In this section, we empirically validate the effectiveness of TOP-D. Specifically, Section 5.1 details our validation datasets, model configurations, baseline methods, and implementation hyperparameters. Section 5.2 presents a comprehensive comparison of TOP-D against existing post-training baselines. Finally, Section 5.3 provides an in-depth analysis of the core components of our method, including the external proximal teacher and the internal trust region iterations, as well as the sensitivity analysis.

5.1 Experimental Setup

Datasets and benchmarks. We conduct our training phase using the DAPO-Math-17k [24] dataset, a high-quality mathematical reasoning corpus. For validation, we benchmark the post-trained models on widely recognized and challenging mathematical reasoning competitions like AIME and AMC.

Models and baselines. To comprehensively investigate the scalability of TOP-D and its robustness across different teacher-student capacity gaps, we employ two distinct base models as students: Qwen3-1.7B-Base [22] and Qwen3-8B-Base [22]. For the target teacher policies, we utilize Qwen3-14B [22] and the exceptionally capable Qwen3-30B-A3B-Instruct-2507 [22]. We compare TOP-D against several post-training paradigms including GRPO [16], DAPO [24], and standard OPD [12].

Implementation details. During the rollout phase, the student model generates 8 responses for each prompt. To execute the trust region iterations, we follow the same off-policy setting as RLVR by dividing the same batch of data into multiple mini-batches, where the global batch size is set to 512 prompts (4096 samples) and the mini-batch size is set to 32 prompts (256 samples). The core interpolation coefficient α for TOP-D is set to 0.1 or 0.2. During training, we apply a sampling temperature of 1.0 and a top-p of 1.0. For performance validation, we use a temperature of 1.0 and a top-p of 0.7. Comprehensive details regarding other hyperparameters are provided in Appendix D.

Student Model	Paradigm	Method	Teacher Model					
			Qwen3-30B-A3B-Instruct-2507					
			avg@32				avg@8	
			AIME24	AIME25	AIME26	AMC23	MATH-500	Olympiad
Qwen3-8B-Base	**Base**	-	9.38	8.02	6.15	53.13	75.23	40.08
	RLVR	GRPO	30.10	22.08	21.67	56.33	76.83	44.92
		DAPO	32.92	27.81	32.29	65.39	81.65	44.23
	Distillation	OPD	24.58	23.33	25.42	76.88	87.98	59.29
		TOP-D (Ours)	**50.42** (+25.84)	**34.06** (+10.73)	**44.06** (+18.64)	**88.13** (+11.25)	**91.23** (+3.25)	**64.67** (+5.38)

Table 2: Performance comparison of the Qwen3-8B-Base across various mathematical benchmarks.

Student Model	Paradigm	Method	Teacher Model					
			Qwen3-30B-A3B-Instruct-2507			*Qwen3-14B*		
			avg@32					
			AIME24	AIME25	AIME26	AIME24	AIME25	AIME26
Qwen3-1.7B-Base	**Base**	-	3.44	1.98	1.88	3.44	1.98	1.88
	RLVR	GRPO	10.52	7.19	4.69	10.52	7.19	4.69
		DAPO	12.29	11.56	8.54	12.29	11.56	8.54
	Distillation	OPD	8.96	7.50	5.94	7.81	8.33	5.63
		TOP-D (Ours)	**20.31** (+11.35)	**17.71** (+10.21)	**13.75** (+7.81)	**12.81** (+5.00)	**14.69** (+6.36)	**9.69** (+4.06)

Table 3: Performance comparison of the Qwen3-1.7B-Base across AIME benchmarks.

5.2 Main Results

Table 2 and Table 3 summarize the validation results of TOP-D compared to other baselines. TOP-D consistently outperforms the evaluated baselines across all mathematical benchmarks and student model scales. The most striking improvement is observed when compared to standard OPD. For instance, using the Qwen3-8B-Base student supervised by the Qwen3-30B-A3B-Instruct-2507 teacher, TOP-D achieves an avg@32 accuracy of 50.42% on AIME24, delivering a massive 25.84% absolute

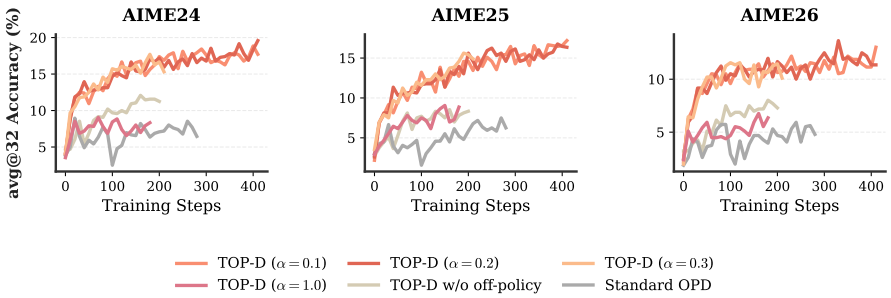

Figure 5: Learning curves for the ablation study on the Qwen3-1.7B-Base. We isolate the effects of the external proximal teacher ($\alpha = 1.0$) and the internal trust region iterations (w/o off-policy).

improvement over standard OPD (24.58%). This immense performance gain definitively proves that bounding the variance of the distillation reward effectively stabilizes the training process and subsequently unlocks the student's reasoning potential. Furthermore, TOP-D firmly dominates the RLVR baselines, exceeding the highly competitive DAPO by 17.5% absolute accuracy on the same benchmark, demonstrating the superiority of stabilized TOP-D over the traditional RLVR baselines.

When scaling down to the smaller Qwen3-1.7B-Base student, standard OPD suffers from severe performance degradation, noticeably falling behind the RLVR baselines. This fragility highlights the vulnerability of the unbounded logarithmic probability ratio when the capacity gap between the teacher and the student is severe. In stark contrast, TOP-D consistently maintains its substantial advantage. For the 1.7B student, TOP-D pushes the AIME24 accuracy past the 20% threshold and consistently surpasses standard OPD by over 10 percentage points on both AIME24 and AIME25.

5.3 Ablation Study

To deeply understand the mechanisms driving TOP-D's success, we isolate the impact of its core structural designs. For this analysis, we utilize the Qwen3-1.7B-Base student model supervised by the Qwen3-30B-A3B-Instruct-2507 teacher. Figure 5 illustrates the learning curves of TOP-D compared to its ablated variants and the standard OPD baseline throughout the entire model training process.

By setting the interpolation coefficient to $\alpha = 1.0$, we remove the external proximal teacher and revert the objective to the unbounded standard OPD reward. As shown in the training curves, this ablation causes highly unstable training dynamics and a significant degradation in overall performance. On the other hand, by disabling the internal trust region iterations (denoted as w/o off-policy), we force the optimization into a strictly on-policy regime. This leads to a drastic drop in sample efficiency, requiring significantly more training steps to converge and climb the performance curve. These empirical results clearly demonstrate that both the *external proximal teacher* and the *internal trust region iterations* are indispensable components for achieving optimization stability and rapid performance improvement. Furthermore, we evaluate the sensitivity of TOP-D to the interpolation coefficient by varying $\alpha \in \{0.1, 0.2, 0.3\}$. We empirically observe that the training stability, convergence speed, and final performance exhibit marginal differences across this interval. This demonstrates that TOP-D is highly robust to the choice of α, functioning as a reliable solution without the need for exhaustive and computationally expensive hyperparameter optimization efforts.

6 Conclusion

In this paper, we introduce Trust Region Policy Distillation (TOP-D), a principled post-training paradigm resolving the optimization instability and sample inefficiency of standard On-Policy Distillation (OPD). Theoretically, TOP-D bounds gradient variance via an external proximal teacher and guarantees monotonic improvement with safe off-policy data reuse through internal trust region iterations. Empirically, it achieves massive performance gains on mathematical reasoning benchmarks, decisively outperforming standard OPD and competitive RLVR baselines across multiple model scales. By achieving these advancements without any additional computational overhead, TOP-D establishes a highly reliable, sample-efficient, and robust paradigm for aligning foundation models.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous reviewers for their comments. This work was supported by the National Natural Science Foundation of China under Grant No. 62506317 and The Hong Kong University of Science and Technology (Guangzhou) Kunpeng&Ascend Center of Cultivation.

References

[1] Changyu Chen, Xiting Wang, and Rui Yan. Extreme region policy distillation. *arXiv preprint arXiv:2605.25582*, 2026.

[2] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[3] Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

[4] Yufei Gu and Zeke Xie. Mano: Restriking manifold optimization for llm training. *arXiv preprint arXiv:2601.23000*, 2026.

[5] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=5h0qf7IBZZ`.

[6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[7] Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. *arXiv preprint arXiv:2601.07155*, 2026.

[8] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

[9] Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

[10] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[11] Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

[12] Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

[13] Daniil Plyusov, Alexey Gorbatovski, Alexey Malakhov, Nikita Balagansky, Boris Shaposhnikov, Daria Korotyshova, and Daniil Gavrilov. Trust-region behavior blending for on-policy distillation. *arXiv preprint arXiv:2605.31159*, 2026.

[14] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[16] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[17] Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

[18] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[19] Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

[20] Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hutter, and Renjing Xu. Simple policy optimization. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=SG8Yx1FyeU`.

[21] Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

[22] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[23] Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, et al. Nemotron-cascade 2: Post-training llms with cascade rl and multi-domain on-policy distillation. *arXiv preprint arXiv:2603.19220*, 2026.

[24] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[25] Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

[26] Ji Zhao, Shitong Shao, Yufei Gu, Xun Zhou, Zeke Xie, et al. Late-to-early training: Let llms learn earlier, so faster and better. In *International Conference on Learning Representations*, volume 2026, pages 98517–98543, 2026.

[27] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

[28] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37: 62557–62583, 2024.

A More Results

Student Model	Paradigm	Method	Teacher Model					
			Qwen3-30B-A3B-Instruct-2507					
			avg@32				avg@8	
			AIME24	AIME25	AIME26	AMC23	MATH-500	Olympiad
Qwen3-1.7B-Base	Base	-	3.44	1.98	1.88	27.27	56.80	24.00
	RLVR	GRPO	10.52	7.19	4.69	35.55	54.73	18.47
		DAPO	12.29	11.56	8.54	37.50	61.88	26.63
	Distillation	OPD	8.96	7.50	5.94	45.70	71.13	35.83
		TOP-D (Ours)	**20.31** (+11.35)	**17.71** (+10.21)	**13.75** (+7.81)	**54.38** (+8.68)	**77.33** (+6.20)	**43.06** (+7.23)

Table 4: Performance comparison of the Qwen3-1.7B-Base across various mathematical benchmarks.

B Related Work

Trust region methods. Schulman et al. [14] proposed Trust Region Policy Optimization (TRPO), which extends the trust region concept to policy gradient methods and provides the first theoretical guarantee of monotonic policy improvement. Proximal Policy Optimization (PPO) [15] overcomes the computationally expensive second-order optimization of TRPO and has been widely adopted. More recently, Simple Policy Optimization (SPO) [20] proposed a quadratic penalty objective function, thereby combining the theoretical guarantees of TRPO with the practical efficiency of PPO smoothly.

Trust region in on-policy distillation. Recent efforts incorporate trust region mechanisms to stabilize OPD against teacher-student distribution mismatches. TrOPD [21] partitions student tokens by teacher acceptance probability, restricting reverse-KL updates to reliable regions while mitigating outlier gradients via forward-KL estimation. TRB [13] constrains the rollout behavior within a student-centered KL ball during early warmup, generating teacher-guided prefixes without altering the per-prefix distillation objective. Meanwhile, ERPD [1] decouples sample and KL efficiency by first extracting supervisory signals through aggressive off-policy optimization, and then distilling them back into the base policy under explicit trust region constraints to filter out spurious policy drift.

Computational Resources

All primary experiments, including the training of the TOP-D and the reproduction of baseline methods, were conducted on a high-performance computing cluster equipped with NVIDIA H200 GPUs. To accelerate the research iteration cycle, our standard distributed training setup utilized 4 compute nodes, with each node containing 8 H200 GPUs (amounting to 32 GPUs in total). However, we emphasize that our proposed method is highly resource-efficient and accessible. The entire TOP-D training pipeline and its reported results can be fully reproduced on a single standard 8-GPU server.

Limitations

Although TOP-D establishes a highly reliable and sample-efficient paradigm for policy distillation, yielding massive improvements over existing baselines, our current validation is naturally bounded by computational budgets and time constraints. Specifically, our primary validations focus on student models up to the 8B parameter scale. Due to resource limitations, we have not yet investigated the scaling behavior of TOP-D on massive-scale foundation models (e.g., scaling the student model beyond 30B parameters). However, given the theoretically grounded variance-reduction properties of our proximal teacher and the guaranteed monotonic improvement from our trust region iterations, we are optimistic that the robust performance of TOP-D will seamlessly transfer to larger-scale regimes.

Furthermore, our current empirical evaluations are constrained to a relatively short optimization window of only approximately 200 to 400 update steps (except for the RLVR baselines). Crucially, we have not observed any empirical signs of performance saturation within this timeframe. This strongly suggests that the impressive results reported in our main text are far from their upper bounds, leaving substantial room for further improvements with continued training. We leave the exploration of such extreme-scale distillation and extended training horizons as exciting avenues for future work.

C Technical Details

In this section, we elaborate on the asymptotic error in Section C.1 and monotonic improvement in Section C.2 to further clarify our main theoretical results.

C.1 Asymptotic Error

While the factor $1/\alpha$ in Theorem 4.4 seemingly suggests that a small α amplifies the asymptotic bound ϵ_∞/α, we show that this bound is actually invariant to α; thus, without internal trust region iterations, TOP-D shares the exact asymptotic optimization error of standard OPD. Specifically, for $\tilde{\pi}_k^* = \alpha\pi^* + (1-\alpha)\pi_k$, the initial target distance from the current student π_k scales linearly with α:

$$d(\pi_k, \tilde{\pi}_k^*) = \mathbb{E}_{x\sim\mathcal{D}_x}\left[\|\pi_k(\cdot \mid x) - \tilde{\pi}_k^*(\cdot \mid x)\|_1\right] = \alpha\, d(\pi_k, \pi^*). \tag{16}$$

Since the proximal target $\tilde{\pi}_k^*$ is α-closer to π_k along the interpolation path toward π^* than in standard OPD ($\alpha = 1$), the single-step optimization residual $\|\epsilon_k\|_1 = d(\pi_{k+1}, \tilde{\pi}_k^*)$ under the same single-step approximation capability scales proportionally with the initial target distance, i.e., $\|\epsilon_k\|_1 = \alpha\|\epsilon_k^{\mathrm{OPD}}\|_1$. Taking the limit supremum as $k \to \infty$ gives $\epsilon_\infty = \alpha\epsilon_\infty^{\mathrm{OPD}}$, which yields

$$\limsup_{k\to\infty} d(\pi_{k+1}, \pi^*) \le \frac{\epsilon_\infty}{\alpha} = \epsilon_\infty^{\mathrm{OPD}}. \tag{17}$$

Consequently, any further effort to reduce $\|\epsilon_k\|_1$ directly tightens the asymptotic bound. Dividing (15) by $\alpha > 0$ gives

$$\frac{\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}_{\mathrm{KL}}^{(0)}\right]}{\alpha} \ge \frac{\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}_{\mathrm{KL}}^{(1)}\right]}{\alpha} \ge \cdots \ge \frac{\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}_{\mathrm{KL}}^{(n)}\right]}{\alpha}. \tag{18}$$

For the output $\pi_{k+1} = \pi_k^{(n)}$, Pinsker's and Jensen's inequalities (Section 4.3) further imply

$$\frac{\|\epsilon_k\|_1}{\alpha} = \frac{d(\pi_k^{(n)}, \tilde{\pi}_k^*)}{\alpha} \le \frac{\mathbb{E}_{x\sim\mathcal{D}_x}\left[\sqrt{2\mathcal{D}_{\mathrm{KL}}^{(n)}}\right]}{\alpha} \le \frac{\sqrt{2\mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathcal{D}_{\mathrm{KL}}^{(n)}\right]}}{\alpha} \le \frac{\delta}{\alpha}, \tag{19}$$

showing that lower expected KL divergence tightens the Pinsker upper bound on $\|\epsilon_k\|_1/\alpha$.

C.2 Practical Objective

Building on the insights of Xie et al. [20], the performance lower bound in Theorem 4.9 can be transformed into a practical form. Specifically, by applying importance sampling and employing the TV divergence identity, we have

$$\begin{aligned}\zeta_\pi(\tilde{\pi}) &= \eta(\pi) + \ell_\pi \mathbb{E}_{s\sim d_{\mathrm{norm}}^\pi(\cdot), a\sim\tilde{\pi}(\cdot|s)}\left[A^\pi(s,a)\right] \\ &= \eta(\pi) + \ell_\pi \mathbb{E}_{s\sim d_{\mathrm{norm}}^\pi(\cdot), a\sim\pi(\cdot|s)}\left[\frac{\tilde{\pi}(a \mid s)}{\pi(a \mid s)} \cdot A^\pi(s,a)\right],\end{aligned} \tag{20}$$

and

$$\mathbb{E}_{s\sim d_{\mathrm{norm}}^\pi(\cdot)}\left[\mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s), \pi(\cdot \mid s))\right] = \frac{1}{2}\mathbb{E}_{s\sim d_{\mathrm{norm}}^\pi(\cdot), a\sim\pi(\cdot|s)}\left[\left|\frac{\tilde{\pi}(a \mid s)}{\pi(a \mid s)} - 1\right|\right]. \tag{21}$$

We now arrive at the following constrained optimization problem:

$$\max_{\tilde{\pi}} \ \mathbb{E}_{s\sim d_{\mathrm{norm}}^\pi(\cdot), a\sim\pi(\cdot|s)}\left[\frac{\tilde{\pi}(a \mid s)}{\pi(a \mid s)} \cdot A^\pi(s,a)\right] \ \text{ s.t. } \ \mathbb{E}_{s\sim d_{\mathrm{norm}}^\pi(\cdot), a\sim\pi(\cdot|s)}\left[\left|\frac{\tilde{\pi}(a \mid s)}{\pi(a \mid s)} - 1\right|\right] \le \epsilon. \tag{22}$$

To efficiently solve this constrained optimization problem using first-order methods, PPO employs a heuristic ratio-clipping mechanism to approximately enforce the TV divergence trust region constraints, yielding the practical objective in (7), which we adapt for our internal trust region iterations.

As described by Xie et al. [20], PPO's hard clipping zeros the gradients of certain data points, leading to a lack of corrective gradients to consistently enforce trust region constraints. Therefore, Simple Policy Optimization (SPO) is designed to make the policy optimization process corrective and stable.

D Hyperparameters

Hyperparameter	GRPO [16]	DAPO [24]	OPD [12]	TOP-D (Ours)
General				
Max prompt length	2048	2048	2048	2048
Optimization				
Global batch size	512	512	512	512
Mini-batch size	32	32	512	32
Mini-batches	16	16	1	16
Group size	8	8	8	8
Off-policy epoch	1	1	-	1
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	1e-6	1e-6	1e-6	1e-6
Warmup steps	10	10	-	-
Sampling				
Rollout max response length	16384	16384	16384	16384
Rollout temperature	1.0	1.0	1.0	1.0
Rollout top-p	1.0	1.0	1.0	1.0
Validation max response length	16384	16384	16384	16384
Validation temperature	1.0	1.0	1.0	1.0
Validation top-p	0.7	0.7	0.7	0.7
Algorithm-specific				
ϵ_{low}	0.2	0.2	-	0.2
ϵ_{high}	0.2	0.28	-	0.2
Interpolation α	-	-	-	0.1 / 0.2

Table 5: Detailed hyperparameters for different methods.

E Proofs

E.1 Proof of Theorem 4.2

Theorem 4.2. *Under Assumption 4.1, for any interpolation coefficient $\alpha \in (0, 1)$, the variance of the TOP-D gradient estimator $\tilde{g}_k$ does not diverge across the entire probability space. Instead, it is strictly bounded by a uniform absolute limit:*

$$\mathrm{tr}(\mathrm{Var}(\tilde{g}_k)) \leq M^2 \left(\log^2(1-\alpha) + C^* \alpha \right), \tag{23}$$

where C^* is a universal mathematical constant independent of the vocabulary size.

Proof. For any square-integrable random vector g, $\mathrm{tr}(\mathrm{Var}(g)) = \mathbb{E}[\|g\|^2] - \|\mathbb{E}[g]\|^2 \leq \mathbb{E}[\|g\|^2]$. Thus, a uniform finite bound on $\mathbb{E}[\|\tilde{g}_k\|^2]$ also establishes the required square integrability.

The TOP-D immediate reward is $\tilde{r}_k = \log(\alpha \rho_k + 1 - \alpha)$. Fix a nonterminal context $(x, y_{<k})$, and write $p_v = \pi_\theta(v \mid x, y_{<k})$ and $q_v = \pi^*(v \mid x, y_{<k})$. For full-vocabulary softmax policies, $p_v, q_v \in (0, 1)$, and we have

$$\mathbb{E}_{y_k \sim \pi_\theta(\cdot|x, y_{<k})} \left[\tilde{r}_k^2 \right] = \sum_{v \in \mathcal{V}} p_v \log^2 \left(\alpha \frac{q_v}{p_v} + 1 - \alpha \right). \tag{24}$$

For one token, write $p = p_v \in (0, 1)$ and $q = q_v \in (0, 1)$. Finite logits give strictly positive softmax probabilities over the full vocabulary. Define

$$h(p, q) = p \log^2 \left(\alpha \frac{q}{p} + 1 - \alpha \right). \tag{25}$$

We split the vocabulary into $\mathcal{V}^- = \{v \in \mathcal{V} : q_v < p_v\}$ and $\mathcal{V}^+ = \{v \in \mathcal{V} : q_v \geq p_v\}$, retaining the probability weights in both regions.

Case 1: The negative reward region $q < p$.

Here $0 < \frac{q}{p} < 1$, so $1 - \alpha < \alpha \frac{q}{p} + 1 - \alpha < 1$. The lower endpoint is approached as $q \to 0^+$ for fixed $p > 0$, and the logarithm is negative throughout this interval. Its squared magnitude is therefore bounded by the endpoint limit, giving

$$\log^2 \left(\alpha \frac{q}{p} + 1 - \alpha \right) \leq \log^2(1 - \alpha). \tag{26}$$

Multiplying by p and summing over the negative region gives

$$\sum_{v \in \mathcal{V}^-} h(p_v, q_v) \leq \log^2(1 - \alpha) \sum_{v \in \mathcal{V}^-} p_v \leq \log^2(1 - \alpha). \tag{27}$$

Case 2: The nonnegative reward region $q \geq p$.

Here $q/p \geq 1$, and hence $1 \leq \alpha q/p + 1 - \alpha \leq \alpha q/p + 1$. Both logarithms are nonnegative, so replacing the inner term by the larger one preserves the squared inequality. Crucially, we retain the teacher probability q:

$$h(p, q) \leq p \log^2 \left(\alpha \frac{q}{p} + 1 \right). \tag{28}$$

Set $u = q/p \in [1, \infty)$, so that $p = q/u$. This substitution separates the teacher's probability mass from the likelihood ratio, rather than bounding each token by an unweighted constant. The factor q must remain until the vocabulary sum is taken. In particular,

$$h(p, q) \leq \frac{q}{u} \log^2(\alpha u + 1) = q\, f(u). \tag{29}$$

The remaining task is to bound the same scalar envelope uniformly over all $u \geq 1$:

$$f(u) = \frac{\log^2(\alpha u + 1)}{u}. \tag{30}$$

To find the global maximum of $f(u)$ on $[1, \infty)$, we compute its derivative with respect to u:

$$\begin{aligned} f'(u) &= \frac{2\alpha \log(1+\alpha u)}{u(1+\alpha u)} - \frac{\log^2(1+\alpha u)}{u^2} \\ &= \frac{\log(1+\alpha u)}{u^2} \left(\frac{2\alpha u}{1+\alpha u} - \log(1+\alpha u) \right). \end{aligned} \tag{31}$$

Setting $f'(u) = 0$ identifies the stationary points. Since $u \geq 1$ and $\alpha \in (0, 1)$, the term $\log(\alpha u+1) > 0$ and can be safely factored out. The condition simplifies to

$$\frac{2\alpha u}{\alpha u + 1} = \log(\alpha u + 1). \tag{32}$$

We further introduce a variable substitution $t = \alpha u + 1$. Because $\alpha u > 0$, we have $t > 1$. Since $\alpha u = t - 1$, the condition transforms into a pure transcendental equation independent of α:

$$2\left(1 - \frac{1}{t}\right) = \log t \implies \log t - 2\left(1 - \frac{1}{t}\right) = 0. \tag{33}$$

Let $w(t) = \log t - 2\left(1 - \frac{1}{t}\right)$. We analyze its roots for $t > 1$:

1. The derivative is $w'(t) = \frac{1}{t} - \frac{2}{t^2} = \frac{t-2}{t^2}$.
2. For $t \in (1, 2)$, $w'(t) < 0$, meaning $w(t)$ strictly decreases. Since $\lim_{t\to 1^+} w(t) = 0$, $w(t) < 0$ for all $t \in (1, 2)$.
3. For $t > 2$, $w'(t) > 0$, meaning $w(t)$ strictly increases.
4. Since $w(2) = \log 2 - 1 < 0$ and $w(e^2) = \frac{2}{e^2} > 0$, by the Intermediate Value Theorem, there exists exactly one unique root t^* in the interval $(2, e^2)$.

Since $f'(u) = -\log(1+\alpha u)w(1+\alpha u)/u^2$, its sign is positive before $u^* = (t^* - 1)/\alpha$ and negative afterwards. Also, $1 + \alpha < 2 < t^*$ ensures $u^* > 1$, so the maximizer lies inside the domain for every admissible α. Substituting this unique global maximizer gives

$$f(u) \leq f(u^*) = \frac{\log^2 t^*}{t^* - 1} \cdot \alpha. \tag{34}$$

Let $C^* = \frac{\log^2 t^*}{t^*-1} = \sup_{z>0} \frac{\log^2(1+z)}{z}$. Then $h(p, q) \leq qf(u) \leq C^* \alpha q$ for $q \geq p$.

Summing the probability-weighted bound over the nonnegative region uses at most the teacher's total probability mass, rather than a separate constant for each vocabulary token:

$$\sum_{v \in \mathcal{V}^+} h(p_v, q_v) \leq C^* \alpha \sum_{v \in \mathcal{V}^+} q_v \leq C^* \alpha. \tag{35}$$

The two regions partition the vocabulary. Combining their bounds controls the whole conditional expectation, regardless of the context $(x, y_{<k})$ or vocabulary size. No multiplication by the number of tokens is needed:

$$\begin{aligned} \mathbb{E}_{y_k \sim \pi_\theta(\cdot|x, y_{<k})} \left[\tilde{r}_k^2\right] &= \sum_{v \in \mathcal{V}^-} h(p_v, q_v) + \sum_{v \in \mathcal{V}^+} h(p_v, q_v) \\ &\leq \log^2(1-\alpha) + C^* \alpha. \end{aligned} \tag{36}$$

Now, we move to the TOP-D gradient estimator $\tilde{g}_k = \nabla_\theta \log \pi_\theta(y_k \mid x, y_{<k}) \cdot \tilde{r}_k$. According to Assumption 4.1, we have

$$\|\tilde{g}_k\|^2 = \|\nabla_\theta \log \pi_\theta(y_k \mid x, y_{<k})\|^2 \cdot \tilde{r}_k^2 \leq M^2 \tilde{r}_k^2. \tag{37}$$

Set the reward and estimator to zero after termination. Averaging the context-uniform second-moment bound over prompts, on-policy prefixes, and tokens gives

$$\begin{aligned} \mathbb{E}\left[\|\tilde{g}_k\|^2\right] &= \mathbb{E}_{x \sim \mathcal{D}_x, y_{<k} \sim \pi_\theta(\cdot|x)} \left[\mathbb{E}_{y_k \sim \pi_\theta(\cdot|x, y_{<k})} \left[\|\tilde{g}_k\|^2\right]\right] \\ &\leq \mathbb{E}_{x \sim \mathcal{D}_x, y_{<k} \sim \pi_\theta(\cdot|x)} \left[M^2 \mathbb{E}_{y_k \sim \pi_\theta(\cdot|x, y_{<k})} \left[\tilde{r}_k^2\right]\right] \\ &\leq \mathbb{E}_{x \sim \mathcal{D}_x, y_{<k} \sim \pi_\theta(\cdot|x)} \left[M^2 \left(\log^2(1-\alpha) + C^* \alpha\right)\right] \\ &= M^2 \left(\log^2(1-\alpha) + C^* \alpha\right). \end{aligned} \tag{38}$$

Finally, we return to our variance inequality established at the beginning of the proof:

$$\mathrm{tr}(\mathrm{Var}(\tilde{g}_k)) \leq \mathbb{E}\left[\|\tilde{g}_k\|^2\right] \leq M^2\left(\log^2(1-\alpha) + C^*\alpha\right). \tag{39}$$

Since M, α, and C^* are all finite, the gradient covariance trace is absolutely bounded. The proof is now complete. □

E.2 Proof of Theorem 4.4

Theorem 4.4. *For any initial student policy $\pi_0 \in \Pi$, consider the iterative update process $\pi_{k+1} = \mathcal{T}(\pi_k) + \epsilon_k$, where ϵ_k accounts for the practical optimization error (e.g., function approximation and optimization noise) at the k-th global step, where $k \geq 0$. Then, the following bound holds:*

$$d(\pi_{k+1}, \pi^*) \leq (1-\alpha)^{k+1} d(\pi_0, \pi^*) + \sum_{i=0}^{k} (1-\alpha)^{k-i} \|\epsilon_i\|_1. \tag{40}$$

Furthermore, assuming the optimization process eventually stabilizes such that the asymptotic error is bounded by $\limsup_{k\to\infty} \|\epsilon_k\|_1 \leq \epsilon_\infty$, we have

$$\limsup_{k\to\infty} d(\pi_{k+1}, \pi^*) \leq \frac{\epsilon_\infty}{\alpha}. \tag{41}$$

Proof. We aim to bound the expected L_1 distance between the student policy π_{k+1} and the optimal target teacher policy π^*.

Step 1: Establishing the single-step recurrence relation.

According to the iterative update process defined in the theorem, for any global step $k \geq 0$, the updated policy is given by

$$\pi_{k+1} = \mathcal{T}(\pi_k) + \epsilon_k, \tag{42}$$

where $\mathcal{T} : \Pi \to \Pi$ is the *proximal teacher operator* defined in (11). By utilizing the triangle inequality of the expected L_1 distance metric $d(\cdot, \cdot)$, we decouple the distance between the updated policy π_{k+1} and the target policy π^* into two components:

$$d(\pi_{k+1}, \pi^*) \leq d(\mathcal{T}(\pi_k), \pi^*) + d(\pi_{k+1}, \mathcal{T}(\pi_k)). \tag{43}$$

We analyze the second term on the right-hand side first. As defined in the main text, the proximal teacher is explicitly given by $\tilde{\pi}_k^* = \mathcal{T}(\pi_k) = \alpha\pi^* + (1-\alpha)\pi_k$. Therefore, the distance between the updated policy π_{k+1} and the proximal teacher $\mathcal{T}(\pi_k)$ precisely corresponds to the expected L_1 norm of the optimization error at step k:

$$d(\pi_{k+1}, \mathcal{T}(\pi_k)) = d(\pi_{k+1}, \tilde{\pi}_k^*) = \|\epsilon_k\|_1. \tag{44}$$

For the first term, we substitute the explicit definition of the proximal operator $\mathcal{T}(\pi_k)$. Utilizing the absolute homogeneity and linearity of the expected L_1 norm over $\mathcal{D}_x$, we obtain

$$\begin{aligned} d(\mathcal{T}(\pi_k), \pi^*) &= \mathbb{E}_{x\sim\mathcal{D}_x}\left[\|\alpha\pi^*(\cdot \mid x) + (1-\alpha)\pi_k(\cdot \mid x) - \pi^*(\cdot \mid x)\|_1\right] \\ &= (1-\alpha)\mathbb{E}_{x\sim\mathcal{D}_x}\left[\|\pi_k(\cdot \mid x) - \pi^*(\cdot \mid x)\|_1\right] \\ &= (1-\alpha) d(\pi_k, \pi^*). \end{aligned} \tag{45}$$

Substituting both decoupled terms back into the triangle inequality (43), we establish the strict single-step recurrence relation for any step:

$$d(\pi_{k+1}, \pi^*) \leq (1-\alpha) d(\pi_k, \pi^*) + \|\epsilon_k\|_1. \tag{46}$$

Step 2: Forward unrolling of the recurrence relation.

To derive the explicit global bound for step $k+1$, we recursively unroll this single-step bound forward, starting from the initial step $k = 0$.

For $k = 0$, the single-step recurrence gives

$$d(\pi_1, \pi^*) \leq (1-\alpha) d(\pi_0, \pi^*) + \|\epsilon_0\|_1. \tag{47}$$

For $k = 1$, we apply the recurrence and substitute the previously established bound for $d(\pi_1, \pi^*)$:

$$\begin{aligned} d(\pi_2, \pi^*) &\leq (1-\alpha) d(\pi_1, \pi^*) + \|\epsilon_1\|_1 \\ &\leq (1-\alpha)^2 d(\pi_0, \pi^*) + (1-\alpha)\|\epsilon_0\|_1 + \|\epsilon_1\|_1. \end{aligned} \tag{48}$$

For $k = 2$, we repeat the process by substituting the bound for $d(\pi_2, \pi^*)$:

$$\begin{aligned} d(\pi_3, \pi^*) &\leq (1-\alpha) d(\pi_2, \pi^*) + \|\epsilon_2\|_1 \\ &\leq (1-\alpha)^3 d(\pi_0, \pi^*) + (1-\alpha)^2\|\epsilon_0\|_1 + (1-\alpha)\|\epsilon_1\|_1 + \|\epsilon_2\|_1. \end{aligned} \tag{49}$$

Through rigorous algebraic expansion by repeating this substitution multiple times, a clear geometric pattern emerges. Accumulating these terms yields the closed-form finite-step global bound:

$$d(\pi_{k+1}, \pi^*) \leq (1-\alpha)^{k+1} d(\pi_0, \pi^*) + \sum_{i=0}^{k} (1-\alpha)^{k-i} \|\epsilon_i\|_1. \tag{50}$$

Step 3: Deriving the asymptotic convergence bound.

To determine the asymptotic behavior, we apply the limit supremum ($\limsup_{k\to\infty}$) to both sides of the inequality:

$$\limsup_{k\to\infty} d(\pi_{k+1}, \pi^*) \leq \limsup_{k\to\infty} \left\{ (1-\alpha)^{k+1} d(\pi_0, \pi^*) + \sum_{i=0}^{k} (1-\alpha)^{k-i} \|\epsilon_i\|_1 \right\}. \tag{51}$$

First, the interpolation coefficient satisfies $\alpha \in (0, 1)$. Consequently, the term associated with the initial distance decays exponentially to zero:

$$\limsup_{k\to\infty} (1-\alpha)^{k+1} d(\pi_0, \pi^*) = 0. \tag{52}$$

Next, we analyze the accumulated error summation term. We are given the assumption that the asymptotic optimization error is bounded, i.e., $\limsup_{k\to\infty} \|\epsilon_k\|_1 \leq \epsilon_\infty$. By the mathematical definition of the limit supremum, for any arbitrarily small constant $\delta > 0$, there exists a sufficiently large integer $N > 0$ such that for all $i \geq N$, the condition $\|\epsilon_i\|_1 \leq \epsilon_\infty + \delta$ strictly holds.

Let $k \to \infty$; we now split the summation into two parts at this critical index N:

$$\sum_{i=0}^{k} (1-\alpha)^{k-i} \|\epsilon_i\|_1 = \sum_{i=0}^{N-1} (1-\alpha)^{k-i} \|\epsilon_i\|_1 + \sum_{i=N}^{k} (1-\alpha)^{k-i} \|\epsilon_i\|_1. \tag{53}$$

As $k \to \infty$, the first term vanishes to zero. For the second term, we safely substitute the uniform bound $\|\epsilon_i\|_1 \leq \epsilon_\infty + \delta$ for all terms where $i \geq N$:

$$\sum_{i=N}^{k} (1-\alpha)^{k-i} \|\epsilon_i\|_1 \leq (\epsilon_\infty + \delta) \sum_{i=N}^{k} (1-\alpha)^{k-i}. \tag{54}$$

We introduce a change of variable by letting $j = k - i$. When $i = k$, $j = 0$. When $i = N$, $j = k - N$. This transforms the summation into a standard geometric series:

$$(\epsilon_\infty + \delta) \sum_{i=N}^{k} (1-\alpha)^{k-i} = (\epsilon_\infty + \delta) \sum_{j=0}^{k-N} (1-\alpha)^{j}. \tag{55}$$

Now we have

$$\limsup_{k\to\infty} \sum_{i=N}^{k} (1-\alpha)^{k-i} \|\epsilon_i\|_1 \leq (\epsilon_\infty + \delta) \lim_{k\to\infty} \sum_{j=0}^{k-N} (1-\alpha)^{j} = (\epsilon_\infty + \delta) \sum_{j=0}^{\infty} (1-\alpha)^{j} = \frac{\epsilon_\infty + \delta}{\alpha}. \tag{56}$$

Because the choice of $\delta > 0$ is arbitrary and can be chosen to be infinitesimally small ($\delta \to 0$), the limit supremum of the entire sequence is strictly bounded by ϵ_∞ / α. Combining this bound with the vanishing first term, we have

$$\limsup_{k\to\infty} d(\pi_{k+1}, \pi^*) \leq \frac{\epsilon_\infty}{\alpha}. \tag{57}$$

The proof is now complete. □

E.3 Proof of Theorem 4.9

Theorem 4.9. *Given Assumption 4.6 and Definition 4.7, for any two policies, $\tilde{\pi}$ and π, the following bound holds:*

$$\eta(\tilde{\pi}) \geq \zeta_\pi(\tilde{\pi}) - 2\xi T_{\max} \ell_\pi \mathbb{E}_{s\sim d^\pi_{\text{norm}}(\cdot)} \left[\mathcal{D}_{\text{TV}}(\tilde{\pi}(\cdot \mid s), \pi(\cdot \mid s))\right], \tag{58}$$

where $\zeta_\pi(\tilde{\pi}) = \eta(\pi) + \ell_\pi \mathbb{E}_{s\sim d^\pi_{\text{norm}}(\cdot), a\sim\tilde{\pi}(\cdot|s)} \left[A^\pi(s,a)\right]$, and $\xi = \max_s |\mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)} \left[A^\pi(s,a)\right]|$.

Proof. To establish the monotonic improvement lower bound, we begin by bounding the absolute error between the true expected return $\eta(\tilde{\pi})$ and our constructed surrogate objective $\zeta_\pi(\tilde{\pi})$.

First, we express the unnormalized performance difference:

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_{s\neq s_\perp} d^{\tilde{\pi}}(s) \sum_{a\in\mathcal{V}} \tilde{\pi}(a \mid s) A^\pi(s,a). \tag{59}$$

$\zeta_\pi(\tilde{\pi})$ is anchored to the base policy's normalized state distribution. By applying Definition 4.7, where $d^\pi(s) = \ell_\pi d^\pi_{\text{norm}}(s)$, we can unroll $\zeta_\pi(\tilde{\pi})$ back into the unnormalized measure space:

$$\zeta_\pi(\tilde{\pi}) = \eta(\pi) + \sum_{s\neq s_\perp} d^{\pi}(s) \sum_{a\in\mathcal{V}} \tilde{\pi}(a \mid s) A^\pi(s,a). \tag{60}$$

We isolate the objective approximation error by subtracting $\zeta_\pi(\tilde{\pi})$ from $\eta(\tilde{\pi})$. By applying Hölder's Inequality, we decouple the global state distribution shift from the local expected advantage drift:

$$|\eta(\tilde{\pi}) - \zeta_\pi(\tilde{\pi})| \leq \left(\sum_{s\neq s_\perp} |d^{\tilde{\pi}}(s) - d^\pi(s)|\right) \cdot \max_s \left|\sum_{a\in\mathcal{V}} \tilde{\pi}(a \mid s) A^\pi(s,a)\right|. \tag{61}$$

The second decoupled term is precisely defined as our constant $\xi = \max_s |\mathbb{E}_{a\sim\tilde{\pi}(\cdot|s)} \left[A^\pi(s,a)\right]|$.

Next, we rigorously bound the L_1 norm of the unnormalized global measure shift. Since the unnormalized measure is the sum of marginal state probabilities over time, we expand this across the maximum generation horizon $T_{\max}$ defined in Assumption 4.6, and apply the triangle inequality:

$$\sum_{s\neq s_\perp} |d^{\tilde{\pi}}(s) - d^\pi(s)| \leq \sum_{t=1}^{T_{\max}} \sum_{s\neq s_\perp} |\mathbb{P}(s_t = s \mid \tilde{\pi}) - \mathbb{P}(s_t = s \mid \pi)|. \tag{62}$$

At the first generation step $t = 1$, the initial state s_1 is strictly defined by the input prompt $x \sim \mathcal{D}_x$ and the special initial token y_0. Because this initial distribution is completely independent of the policy, the marginal probability distance is exactly zero:

$$\sum_{s\neq s_\perp} |\mathbb{P}(s_1 = s \mid \tilde{\pi}) - \mathbb{P}(s_1 = s \mid \pi)| = 0. \tag{63}$$

For any subsequent generation step $t > 1$, we construct a telescoping sum to bound the difference in marginal probabilities. Let $\mathbb{P}_i$ denote a *hybrid trajectory distribution* where the first i steps are generated by the behavior policy π, and the remaining steps from $i+1$ to $t-1$ are generated by the target policy $\tilde{\pi}$. Note that $\mathbb{P}_{t-1}$ represents full generation under π, and $\mathbb{P}_0$ represents full generation under $\tilde{\pi}$. The telescoping sum bridging these two extreme distributions is given by

$$\mathbb{P}(s_t = s \mid \tilde{\pi}) - \mathbb{P}(s_t = s \mid \pi) = \sum_{i=1}^{t-1} \left(\mathbb{P}_{i-1}(s_t = s) - \mathbb{P}_i(s_t = s)\right). \tag{64}$$

Taking the absolute value and applying the triangle inequality yields

$$\sum_{s\neq s_\perp} |\mathbb{P}(s_t = s \mid \tilde{\pi}) - \mathbb{P}(s_t = s \mid \pi)| \leq \sum_{i=1}^{t-1} \sum_{s\neq s_\perp} |\mathbb{P}_i(s_t = s) - \mathbb{P}_{i-1}(s_t = s)|. \tag{65}$$

Observe that the adjacent hybrid distributions $\mathbb{P}_i$ and $\mathbb{P}_{i-1}$ share the identical generation history up to step $i-1$ (governed by π) and share the identical rollout strategy from step $i+1$ onward (governed by $\tilde{\pi}$). They diverge only at the specific decision step i.

In our autoregressive generation framework, the environmental transitions are purely deterministic (appending the generated token to the prefix). Consequently, summing over all possible future states s_t effectively marginalizes them out, collapsing the trajectory distance precisely to the expected L_1 divergence of the action distributions at the singular branching step i:

$$\sum_{s \neq s_\perp} |\mathbb{P}_i(s_t = s) - \mathbb{P}_{i-1}(s_t = s)| \leq \mathbb{E}_{s_i \sim \mathbb{P}(s_i \in \cdot | \pi)} \left[\|\tilde{\pi}(\cdot \mid s_i) - \pi(\cdot \mid s_i)\|_1 \right]. \tag{66}$$

Substituting this into our bound and summing over the entire telescoping sequence from $i = 1$ to $t - 1$, we obtain the marginal probability distance for a specific step t:

$$\sum_{s \neq s_\perp} |\mathbb{P}(s_t = s \mid \tilde{\pi}) - \mathbb{P}(s_t = s \mid \pi)| \leq 2 \sum_{i=1}^{t-1} \mathbb{E}_{s_i \sim \mathbb{P}(s_i \in \cdot | \pi)} \left[\mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s_i), \pi(\cdot \mid s_i)) \right]. \tag{67}$$

As established earlier, the initial state distribution at $t = 1$ is strictly fixed by the prompt and the special initial token, yielding zero divergence. Therefore, we sum this temporal bound over the valid response steps from $t = 2$ to the maximum generation horizon $T_{\max}$ to measure the global shift:

$$\sum_{s \neq s_\perp} |d^{\tilde{\pi}}(s) - d^{\pi}(s)| \leq 2 \sum_{t=2}^{T_{\max}} \sum_{i=1}^{t-1} \mathbb{E}_{s_i \sim \mathbb{P}(s_i \in \cdot | \pi)} \left[\mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s_i), \pi(\cdot \mid s_i)) \right]. \tag{68}$$

To efficiently decouple this complex temporal accumulation, we rearrange the nested summations. By exchanging the order of summation between the time steps, we observe that the expected divergence at any step i is repeatedly accumulated exactly $(T_{\max} - i)$ times:

$$\begin{aligned}
&\sum_{t=2}^{T_{\max}} \sum_{i=1}^{t-1} \mathbb{E}_{s_i \sim \mathbb{P}(s_i \in \cdot | \pi)} \left[\mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s_i), \pi(\cdot \mid s_i)) \right] \\
=& \sum_{i=1}^{T_{\max}-1} (T_{\max} - i) \mathbb{E}_{s_i \sim \mathbb{P}(s_i \in \cdot | \pi)} \left[\mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s_i), \pi(\cdot \mid s_i)) \right] \\
\leq& T_{\max} \sum_{i=1}^{T_{\max}} \sum_{s \neq s_\perp} \mathbb{P}(s_i = s \mid \pi) \mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s), \pi(\cdot \mid s)) \\
=& T_{\max} \sum_{s \neq s_\perp} \mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s), \pi(\cdot \mid s)) \left(\sum_{i=1}^{T_{\max}} \mathbb{P}(s_i = s \mid \pi) \right).
\end{aligned} \tag{69}$$

Here, we relax this sum by bounding the varying coefficient $(T_{\max} - i)$ with the uniform maximum horizon $T_{\max}$ and expand the expectation into a summation over specific states.

Since the unnormalized state measure $d^{\pi}(s)$ intrinsically aggregates the expected distribution across all steps by definition, i.e., $d^{\pi}(s) = \sum_{t=1}^{\infty} \mathbb{P}(s_t = s \mid \pi) \geq \sum_{i=1}^{T_{\max}} \mathbb{P}(s_i = s \mid \pi)$, we can concisely upper-bound the relaxed sequence, so

$$\sum_{s \neq s_\perp} |d^{\tilde{\pi}}(s) - d^{\pi}(s)| \leq 2 T_{\max} \sum_{s \neq s_\perp} d^{\pi}(s) \mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s), \pi(\cdot \mid s)). \tag{70}$$

Recalling Definition 4.7, we substitute the unnormalized measure with $d^{\pi}(s) = \ell_\pi d^{\pi}_{\mathrm{norm}}(s)$, gracefully transforming the summation into a formal expectation over the normalized state distribution:

$$\sum_{s \neq s_\perp} |d^{\tilde{\pi}}(s) - d^{\pi}(s)| \leq 2 T_{\max} \ell_\pi \mathbb{E}_{s \sim d^{\pi}_{\mathrm{norm}}(\cdot)} \left[\mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s), \pi(\cdot \mid s)) \right]. \tag{71}$$

Finally, substituting this global measure shift bound back into our decoupled objective approximation error (61), we obtain the absolute bound for the objective difference:

$$|\eta(\tilde{\pi}) - \zeta_\pi(\tilde{\pi})| \leq 2 \xi T_{\max} \ell_\pi \mathbb{E}_{s \sim d^{\pi}_{\mathrm{norm}}(\cdot)} \left[\mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s), \pi(\cdot \mid s)) \right], \tag{72}$$

which implies that

$$\eta(\tilde{\pi}) \geq \zeta_\pi(\tilde{\pi}) - 2 \xi T_{\max} \ell_\pi \mathbb{E}_{s \sim d^{\pi}_{\mathrm{norm}}(\cdot)} \left[\mathcal{D}_{\mathrm{TV}}(\tilde{\pi}(\cdot \mid s), \pi(\cdot \mid s)) \right]. \tag{73}$$

The proof is now complete. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, and the type of optimizer) necessary to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report appropriate information about the statistical significance of the experiments, including the avg@32 accuracy.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources (type of compute workers, memory, and time of execution) required to reproduce each experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the creators or original owners of all assets (e.g., code, data, models) used in the paper, and we explicitly mention and respect their licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.