FGRPO: FEDERATED GRPO WITH ADAPTIVE AGGREGATION ON NON-IID DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in language models have established reinforcement learning as the primary paradigm for eliciting self-correction and long-chain reasoning. While *group relative policy optimization* (GRPO) offers superior scalability by eliminating the critic network, deploying it on a central infrastructure entails collecting a large volume of data from distributed owners, which poses communication overhead and significant privacy risks. To address these concerns, we introduce *federated GRPO* (FGRPO), a framework designed to decentralize the fine-tuning of reasoning models across different data owners. To account for differences in local learning progress under heterogeneous client data, FGRPO incorporates an adaptive aggregation mechanism based on relative performance gain. By characterizing each client's improvement relative to its historical baseline, the framework dynamically prioritizes effective learning trajectories regardless of local data heterogeneity. Rigorous theoretical analysis and extensive experiments are conducted to verify the efficacy of FGRPO. Our code is available at /r/Anonymous-FGRPO/.

1 INTRODUCTION

The paradigm of *large language models* (LLMs) has recently been redefined by the emergence of advanced reasoning frameworks, such as OpenAI-o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), and Qwen2.5-VL-3B-Instruct (Bai et al., 2025b). A key driver behind these advances is the strategic use of *reinforcement learning* (RL), which has emerged as an effective paradigm for eliciting self-correction and robust long-chain reasoning in LLMs (Zhang et al., 2025).

Existing RL methods, such as proximal policy optimization (PPO) (Schulman et al., 2017) and REINFORCE-style variants such as leave-one-out (RLOO) (Ahmadian et al., 2024), estimate advantages from absolute reward signals, typically using value-function critics or variance-reduction baselines. In contrast, *group relative policy optimization* (GRPO) (Shao et al., 2024) leverages intra-group comparisons to compute relative advantages, thereby avoiding the need for a separate critic network. Recent studies (Zheng et al., 2025; Chen et al., 2025b; Zhang & Zuo, 2025; Yu et al., 2025) further improve the robustness and efficiency of GRPO.

Meanwhile, data privacy remains a critical concern in training large-scale reasoning models, as high-quality reasoning traces often contain sensitive intellectual property. Aggregating such data from multiple owners into a centralized third-party infrastructure creates significant risks of privacy leakage and legal non-compliance. To this end, *federated learning* (FL) has emerged as an effective framework for distributed model training, in which a learning task is collaboratively solved by a federation of participating data owners (or clients) (McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020). Specifically, clients communicate only their local updates to the central server for aggregation and do not share their local data with others for the sake of privacy preservation.

Although a substantial body of work has explored *federated reinforcement learning* (FedRL) (Fan et al., 2021; Jin et al., 2022; Khodadadian et al., 2022; Zhang et al., 2024; Woo et al., 2024; Fang et al., 2025; Jiang et al., 2025; Wang et al., 2024; Xiong et al., 2025; Lan et al., 2025), applying FL to LLM fine-tuning has only recently emerged in the form of federated *reinforcement learning with human feedback* (RLHF) (Fan et al., 2025), where human feedback is utilized to define client-specific rewards. In this setting, adopting traditional RL methods such as Actor–Critic (Konda & Tsitsiklis, 1999) and PPO (Schulman et al., 2017) leads to federated variants relying on value (critic) networks and absolute reward signals for advantage estimation. In contrast, as mentioned above, GRPO

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

eliminates the critic and derives advantages from relative comparisons among multiple sampled outputs; this design renders advantages inherently local. Particularly, under non-IID data, where each client's dataset comprises heterogeneous tasks, e.g., ones of varying difficulty, these advantages become incomparable: modest gains on simpler tasks may yield gradients comparable in magnitude to substantial improvements on more complex tasks, leading to gradient interference during aggregation. While prior FedRL approaches address data heterogeneity through shared value estimation (Yang et al., 2024; Xie & Song, 2025), such mechanisms are not applicable for critic-free GRPO.

In this paper, we propose FGRPO, an FL framework that leverages GRPO to fine-tune reasoning-capable LLMs under non-IID data. To compensate for the absence of a global or shared critic, FGRPO introduces *relative performance gain* (RPG) as a decentralized surrogate, shifting aggregation from absolute performance to relative learning progress. Specifically, each client measures its improvement relative to an *exponential moving average* (EMA) baseline and normalizes it by local reward volatility, yielding a progress-aware signal enabling cross-client comparability while suppressing noisy or stagnant updates. The server then aggregates local models using an RPG-based weighting scheme that prioritizes clients demonstrating consistent and meaningful progress relative to their task difficulty. In this way, FGRPO amplifies informative learning trajectories and enables robust optimization under non-IID data. Our main contributions are summarized as follows:

- We propose FGRPO, a framework that extends GRPO to FL settings through an RPG-based adaptive aggregation mechanism that accounts for heterogeneity in local learning progress under non-IID data.
- We provide a rigorous theoretical analysis to characterize the convergence behavior of FGRPO. Specifically, we derive a finite-time bound on the average expected squared gradient norm by an $\mathcal{O}(T^{-1/2})$ transient term and an $\mathcal{O}(1)$ residual term.
- We conduct extensive experiments on standard benchmarks with various models. The results show that FGRPO consistently outperforms state-of-the-art baselines, demonstrating superior scalability and robustness across varying client populations and non-IID settings.

The remainder of this paper is organized as follows. Sec. 2 introduces the system model and preliminaries. The detailed design of our FGRPO framework and the corresponding analysis are provided in Sec. 3. We report extensive experiment results in Sec. 4 and survey the related literature in Sec. 5. Finally, Sec. 6 concludes the paper.

2 System Model and Preliminaries

2.1 Reinforcement Learning and GRPO

Consider an autoregressive language model with a finite token vocabulary, defined as a policy π_θ parameterized by θ. Let q denote a prompt and o denote an output sequence. $\pi_\theta(o_\ell \mid q, o_{<\ell})$ is the conditional probability of generating output token o_ℓ given prompt q and previous outputs $o_{<\ell} = (o_1, o_2, \ldots, o_{\ell-1})$. Hence, $\pi_\theta(o \mid q) = \prod_{\ell=1}^{|o|} \pi_\theta(o_\ell \mid q, o_{<\ell})$, where $|o|$ represents the length of the output sequence o. We define a reward function $r(q, o)$, e.g., a rule-based verifier, to measure the quality of the output o given prompt q. Let $r_{\min} < r_{\max}$ be finite reward bounds such that $r_{\min} \le r(q, o) \le r_{\max}$ for every relevant prompt and response. The goal is to optimize the model parameters θ to maximize the expected reward over a distribution of prompts $\mathcal{D}$. Formally, the objective is to solve the following optimization problem:

$$\max_{\theta \in \mathbb{R}^d} \mathcal{J}(\theta) \triangleq \mathbb{E}_{q \sim \mathcal{D}, o \sim \pi_\theta(\cdot|q)}[r(q, o)]. \tag{1}$$

Responses stop at the first end-of-sequence token or at a fixed length cap $\ell_{\max} \ge 1$. A sampled end-of-sequence token is included in o, so $1 \le |o| \le \ell_{\max}$. Generally, the above problem can be resolved by gradient ascent:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{q \sim \mathcal{D},\, o \sim \pi_\theta(\cdot|q)} \left[r(q, o) \nabla_\theta \log \pi_\theta(o \mid q) \right]. \tag{2}$$

Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025) uses a surrogate objective with group-relative advantages and KL regularization. For each prompt $q \sim \mathcal{D}$, independently sample $K \ge 2$ responses from the frozen behavior policy, $o_k \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$ for $k = 1, \ldots, K$.

The parameter θ_{old} remains fixed while the current parameter θ is updated using these responses. Let $r_k = r(q, o_k)$, and define the group relative advantage estimation as

$$A_k = \frac{r_k - \text{mean}(r_1, \ldots, r_K)}{\text{std}(r_1, \ldots, r_K)}.$$

For each token position $\ell = 1, 2, \ldots, |o_k|$ within the output o_k, let

$$\begin{aligned} &\Phi_{\theta,\theta_{\text{old}}}\left(o_{k,\ell} \mid q, o_{k,<\ell}\right) \\ &\triangleq \min\left\{\rho_{\theta,\theta_{\text{old}}}(o_{k,\ell} \mid q, o_{k,<\ell})A_k, \text{clip}(\rho_{\theta,\theta_{\text{old}}}(o_{k,\ell} \mid q, o_{k,<\ell}), 1-\varepsilon_{\text{clip}}, 1+\varepsilon_{\text{clip}})\, A_k\right\}, \end{aligned} \tag{3}$$

where

$$\rho_{\theta,\theta_{\text{old}}}(o_{k,\ell} \mid q, o_{k,<\ell}) = \pi_\theta(o_{k,\ell} \mid q, o_{k,<\ell}) / \pi_{\theta_{\text{old}}}(o_{k,\ell} \mid q, o_{k,<\ell}), \tag{4}$$

denotes the token-level probability ratio between the current policy π_θ and old one $\pi_{\theta_{\text{old}}}$, and $\varepsilon_{\text{clip}} > 0$ is the clipping parameter. GRPO also penalizes deviations from a fixed reference policy π_{ref} with parameters θ_{ref}. For a sampled token $o_{k,\ell}$, define the nonnegative sampled KL penalty

$$\widehat{\mathbb{D}}_{k,\ell}(\theta; q) \triangleq \frac{\pi_{\text{ref}}(o_{k,\ell} \mid q, o_{k,<\ell})}{\pi_\theta(o_{k,\ell} \mid q, o_{k,<\ell})} - \log \frac{\pi_{\text{ref}}(o_{k,\ell} \mid q, o_{k,<\ell})}{\pi_\theta(o_{k,\ell} \mid q, o_{k,<\ell})} - 1. \tag{5}$$

The GRPO surrogate objective is to maximize

$$\mathcal{J}_{\text{GRPO}}(\theta) \triangleq \mathbb{E}_{q\sim\mathcal{D},\{o_k\}_{k=1}^K \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{K} \sum_{k=1}^{K} \frac{1}{|o_k|} \sum_{\ell=1}^{|o_k|} \left(\Phi_{\theta,\theta_{\text{old}}}\left(o_{k,\ell} \mid q, o_{k,<\ell}\right) - \beta \widehat{\mathbb{D}}_{k,\ell}(\theta; q) \right) \right], \tag{6}$$

where π_{ref} denotes a reference policy, and $\beta \geq 0$ is a hyperparameter controlling the KL penalty.

2.2 Federated Reinforcement Learning

Consider a Federated Learning (FL) setting with clients $\mathcal{N} = \{1, 2, \ldots, N\}$, where each client $i \in \mathcal{N}$ has its own local dataset $\mathcal{D}_i$ of prompts. The FL setting also involves a central server that coordinates the training process across all clients. The goal is to collaboratively train a global model parameterized by θ across all clients while retaining raw prompts on the clients. Formally, the global objective in the FL setting can be expressed as

$$\min_{\theta \in \mathbb{R}^d} F(\theta) \triangleq \sum_{i=1}^{N} p_i F_i(\theta), \text{ where } F_i(\theta) \triangleq \mathbb{E}_{\zeta\sim\mathcal{D}_i}[f_i(\theta; \zeta)]. \tag{7}$$

The function f_i represents the local loss function of client i. The weights $p_i > 0$ satisfy $\sum_{i=1}^N p_i = 1$ and represent the contributions of the clients to the target objective; write $p_{\min} \triangleq \min_i p_i > 0$. Typically, p_i is proportional to the size of client i's dataset, i.e., $p_i = |\mathcal{D}_i| / \sum_{j=1}^N |\mathcal{D}_j|$.

By applying the RL paradigm to the FL setting, the objective can be formulated as

$$\max_{\theta \in \mathbb{R}^d} \quad \mathcal{J}(\theta) \triangleq \sum_{i=1}^{N} p_i \mathcal{J}_i(\theta), \tag{8}$$

where $\mathcal{J}_i(\theta) \triangleq \mathbb{E}_{q\sim\mathcal{D}_i, o\sim\pi_\theta(\cdot|q)}[r(q, o)]$ denotes the local objective function of client i. Designing a federated GRPO framework on non-IID data is particularly challenging because GRPO is critic-free and derives advantages through intra-group normalization, making the resulting learning signals inherently local and not directly comparable across clients. In heterogeneous settings, clients may face prompts with very different difficulty levels and reward distributions, so the same normalized improvement can correspond to very different absolute progress across clients. Without a global or shared critic to anchor these signals, standard federated averaging becomes unreliable and may aggregate misaligned updates, leading to gradient mismatch and unstable optimization. Consequently, the core challenge is to restore cross-client comparability while remaining fully critic-free.

3 FGRPO

3.1 Overview

In each communication round $t = 0, 1, \ldots, T-1$, the central server broadcasts the current global model parameters $\theta^{[t]}$ to all clients. Each client i performs E rollout cycles, each followed by U updates on the sampled block. In cycle $e = 0, 1, \ldots, E-1$, for each sampled prompt $q \sim \mathcal{D}_i$, client i generates K responses $\{o_k\}_{k=1}^K$ according to the current policy $\pi_{\theta_i^{[t,e]}}$, evaluates the corresponding rewards $r_{i,k}^{[t,e]}(q) = r(q, o_k)$, and computes the group-relative advantages $A_{i,k}^{[t,e]}(q)$ for $k = 1, 2, \ldots, K$. It then performs U ascent updates on the sampled approximation of the surrogate:

$$\mathcal{F}_i(\theta) \triangleq \mathbb{E}_{q\sim\mathcal{D}_i, \{o_k\}_{k=1}^K \sim \pi_{\theta_i^{[t,e]}}(\cdot|q)} \left[\frac{1}{K} \sum_{k=1}^{K} \frac{1}{|o_k|} \sum_{\ell=1}^{|o_k|} \left(\Phi_{\theta, \theta_i^{[t,e]}}(o_{k,\ell} \mid q, o_{k,<\ell}) - \beta \widehat{\mathbb{D}}_{k,\ell}(\theta; q) \right) \right]$$

Here the behavior policy and the sampled advantages are held fixed during the U reuse updates. After all E cycles, the client uploads its local model parameters $\theta_i^{[t+1]}$ and RPG value $h_i^{[t]}$ to the central server. The server aggregates the local model parameters $\{\theta_i^{[t+1]}\}_{i=1}^N$ with RPG-based weights to calculate a new global model parameter $\theta^{[t+1]}$ in an Adam-like manner.

3.2 Algorithm Description

The pseudocode of our FGRPO algorithm is presented in Algorithm 1. In each communication round $t = 0, 1, \ldots, T-1$, the central server broadcasts the current global model parameters $\theta^{[t]}$ to all clients. Each client i then performs E local rollout-and-update cycles to compute its local update $\theta_i^{[t+1]}$ and RPG value $h_i^{[t]}$. Specifically, at each cycle $e = 0, 1, \ldots, E-1$, client i samples a minibatch of B prompts $\mathcal{B}_i^{[t,e]}$ from its local dataset $\mathcal{D}_i$. For each prompt $q \in \mathcal{B}_i^{[t,e]}$, the client generates K responses $\{o_k\}_{k=1}^K \sim \pi_{\theta_i^{[t,e]}}(\cdot \mid q)$, and obtains reward $r_{i,k}^{[t,e]}(q) = r(q, o_k), \forall k = 1, 2, \ldots, K$. Let $\bar{r}_i^{[t,e]} = \frac{1}{BK} \sum_{q\in\mathcal{B}_i^{[t,e]}} \sum_{k=1}^K r_{i,k}^{[t,e]}(q)$ denote the average reward of client i at local rollout cycle e. Based on the rewards, client i computes its local group-relative advantages

$$A_{i,k}^{[t,e]}(q) \triangleq \frac{r_{i,k}^{[t,e]}(q) - \bar{r}_i^{[t,e]}(q)}{\sqrt{K^{-1} \sum_{j=1}^K (r_{i,j}^{[t,e]}(q) - \bar{r}_i^{[t,e]}(q))^2 + \varepsilon_A^2}}, \text{ where } \bar{r}_i^{[t,e]}(q) \triangleq \frac{1}{K} \sum_{j=1}^{K} r_{i,j}^{[t,e]}(q), \tag{9}$$

where $\varepsilon_A > 0$ is a small stabilization constant. We define the local surrogate objective for client i as

$$\widehat{\mathcal{F}}_i^{[t,e]}(\theta; \theta_i^{[t,e]}) \triangleq \frac{1}{BK} \sum_{q\in\mathcal{B}_i^{[t,e]}} \sum_{k=1}^{K} \frac{1}{|o_k|} \sum_{\ell=1}^{|o_k|} \left(\Phi_{\theta, \theta_i^{[t,e]}}(o_{k,\ell} \mid q, o_{k,<\ell}) - \beta \widehat{\mathbb{D}}_{k,\ell}(\theta; q) \right). \tag{10}$$

The minibatch is an indexed collection of B independent prompt draws. Sums over $q \in \mathcal{B}_i^{[t,e]}$ count each occurrence separately, and responses are sampled independently for each occurrence, even when prompt values coincide. Client i initializes $\theta_i^{[t,e,0]} = \theta_i^{[t,e]}$ and performs U updates:

$$\theta_i^{[t,e,u+1]} = \theta_i^{[t,e,u]} + \hat{\eta} g_i^{[t,e,u]}, \quad \forall u = 0, 1, \ldots, U-1, \tag{11}$$

where the local update direction is

$$g_i^{[t,e,u]} = \frac{1}{BK} \sum_{q\in\mathcal{B}_i^{[t,e]}} \sum_{k=1}^{K} \frac{1}{|o_k|} \sum_{\ell=1}^{|o_k|} \left(\nabla_\theta \Phi_{\theta, \theta_i^{[t,e]}}(o_{k,\ell} \mid q, o_{k,<\ell}) - \beta \nabla_\theta \widehat{\mathbb{D}}_{k,\ell}(\theta; q) \right) \Big|_{\theta=\theta_i^{[t,e,u]}}. \tag{12}$$

Here $\hat{\eta} > 0$ is the local learning rate shared by all clients. Let $\theta_i^{[t,e+1]} = \theta_i^{[t,e,U]}$ denote the local model parameters after the U iterations at rollout cycle e. After the E rollout cycles, client i averages

Algorithm 1 Federated Group Relative Policy Optimization

Require: Initial policy $\theta^{[0]}$, local data $\{\mathcal{D}_i\}_{i=1}^N$, reference policy π_{ref}, and reward r
Require: GRPO parameters $B, K, \varepsilon_{\text{clip}}, \varepsilon_A, \beta$, rollout cycles E, and reuse steps U
Require: Server and client learning rates $\eta, \hat{\eta}$ and communication rounds T
Require: RPG parameters $\lambda_{\text{base}}, \lambda_{\text{clip}}, \sigma_{\min}, \sigma_{\max}, \varepsilon_{\text{rpg}}, \tau_{\min}, \tau_{\max}, \lambda_{\text{anneal}}$
Require: Server parameters $\beta_1, \beta_2, \varepsilon_s$ and initial states $\{\varphi_i^{[-1]}, \sigma_i^{[-1]}\}_{i=1}^N$, $m^{[-1]} = v^{[-1]} = 0$
Ensure: Optimized policy parameters $\theta^{[T]}$

for $t = 0, 1, \ldots, T-1$ **do**
  Server side:
  Broadcast global parameters $\theta^{[t]}$ to clients $\mathcal{N}$
  Client side:
  for each client $i \in \mathcal{N}$ **in parallel do**
    $\theta_i^{[t,0]} = \theta^{[t]}$
    for $e = 0, 1, \ldots, E-1$ **do**
      Sample $q_{i,1}, \ldots, q_{i,B} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_i$ and set $\mathcal{B}_i^{[t,e]} \leftarrow (q_{i,b})_{b=1}^B$
      for all $q \in \mathcal{B}_i^{[t,e]}$ **do**
        Generate K outputs $o_k \sim \pi_{\theta_i^{[t,e]}}(\cdot|q), \forall k = 1, 2, \ldots, K$
        Obtain rewards $r_{i,k}^{[t,e]}(q) = r(q, o_k), \forall k = 1, 2, \ldots, K$
        Compute advantages $A_{i,k}^{[t,e]}(q), \forall k = 1, 2, \ldots, K$, using (9)
      end for
      $\bar{r}_i^{[t,e]} = \frac{1}{BK} \sum_{q \in \mathcal{B}_i^{[t,e]}} \sum_{k=1}^K r_{i,k}^{[t,e]}(q)$
      $\theta_i^{[t,e,0]} = \theta_i^{[t,e]}$
      for $u = 0, 1, \ldots, U-1$ **do**
        Compute $g_i^{[t,e,u]}$ using (12)
        $\theta_i^{[t,e,u+1]} = \theta_i^{[t,e,u]} + \hat{\eta} g_i^{[t,e,u]}$
      end for
      $\theta_i^{[t,e+1]} = \theta_i^{[t,e,U]}$
    end for
    $\theta_i^{[t+1]} = \theta_i^{[t,E]}$
    Calculate $\bar{r}_i^{[t]} = \frac{1}{E} \sum_{e=0}^{E-1} \bar{r}_i^{[t,e]}$
    $\delta_i^{[t]} \leftarrow \bar{r}_i^{[t]} - \varphi_i^{[t-1]}$
    $\sigma_i^{[t]} \leftarrow \text{clip}\Big(\lambda_{\text{clip}} \sigma_i^{[t-1]} + (1 - \lambda_{\text{clip}}) |\delta_i^{[t]}|, \sigma_{\min}, \sigma_{\max}\Big)$
    $h_i^{[t]} \leftarrow \delta_i^{[t]} / (\sigma_i^{[t]} + \varepsilon_{\text{rpg}})$
    $\varphi_i^{[t]} \leftarrow \lambda_{\text{base}} \bar{r}_i^{[t]} + (1 - \lambda_{\text{base}}) \varphi_i^{[t-1]}$
    Upload $\theta_i^{[t+1]}$ and $h_i^{[t]}$ to the server
  end for
  Server side:
  Calculate weight $\omega_i^{[t]}$ for each client $i \in \mathcal{N}$ using Eq. (15)
  Calculate moments $m^{[t]}$ and $v^{[t]}$ using Eqs. (16)–(17)
  Update global model $\theta^{[t+1]} = \theta^{[t]} + \eta \frac{m^{[t]}}{\sqrt{v^{[t]}} + \varepsilon_s}$
end for
return $\theta^{[T]}$

the rewards collected across those cycles: $\bar{r}_i^{[t]} = \frac{1}{E} \sum_{e=0}^{E-1} \bar{r}_i^{[t,e]}$. Let $\theta_i^{[t+1]} = \theta_i^{[t,E]}$ denote the local model returned by client i in round t.

At each round t, each client i also calculates its relative performance gain $h_i^{[t]}$. Specifically, let

$$\varphi_i^{[t]} = \lambda_{\text{base}} \bar{r}_i^{[t]} + (1 - \lambda_{\text{base}}) \varphi_i^{[t-1]} \tag{13}$$

denote the exponential moving average (EMA) of the average reward $\bar{r}_i^{[t]}$ with a base weight $\lambda_{\text{base}} \in (0, 1)$. For every client, initialize $\varphi_i^{[-1]} \in [r_{\min}, r_{\max}]$; the reward bounds and this choice keep the EMA in the same interval. The relative performance gain $h_i^{[t]}$ is then defined as

$$h_i^{[t]} = (\bar{r}_i^{[t]} - \varphi_i^{[t-1]}) \big/ (\sigma_i^{[t]} + \varepsilon_{\text{rpg}}). \tag{14}$$

where $\sigma_i^{[t]} = \text{clip}\left(\lambda_{\text{clip}}\sigma_i^{[t-1]} + (1-\lambda_{\text{clip}})|\bar{r}_i^{[t]} - \varphi_i^{[t-1]}|, \sigma_{\min}, \sigma_{\max}\right)$ is a clipped estimate of the reward fluctuation with $0 < \lambda_{\text{clip}} < 1$. Each client initializes $\sigma_i^{[-1]} \in [\sigma_{\min}, \sigma_{\max}]$. The stabilizer $\varepsilon_{\text{rpg}} > 0$ and the positive volatility floor keep the denominator positive. At the end of each communication round t, each client i sends $\theta_i^{[t+1]}$ and $h_i^{[t]}$ to the central server.

After receiving $\theta_i^{[t+1]}$ and $h_i^{[t]}$ from every client, the server calculates the aggregation weight $\omega_i^{[t]}$ using a Boltzmann distribution with a dynamic temperature $\tau^{[t]}$:

$$\omega_i^{[t]} = \exp\left(h_i^{[t]}/\tau^{[t]}\right) \Big/ \sum_{i'=1}^{N} \exp\left(h_{i'}^{[t]}/\tau^{[t]}\right). \tag{15}$$

where $\tau^{[t]} = \tau_{\min} + (\tau_{\max} - \tau_{\min})\exp(-\lambda_{\text{anneal}} \cdot t)$. Here, the temperature $\tau^{[t]}$ controls the concentration of the aggregation weights, with $0 < \tau_{\min} \leq \tau_{\max}$ denoting its lower and upper bounds and $\lambda_{\text{anneal}} \geq 0$ controlling the annealing rate. The case $\lambda_{\text{anneal}} = 0$ gives a constant temperature. The positive lower bound keeps the softmax well defined. A higher temperature produces more uniform weights, while a lower temperature places greater weight on clients with larger RPG scores. According to the RPG-based weights, the server calculates a global pseudo-gradient

$$\Delta^{[t]} = \sum_{i=1}^{N} \omega_i^{[t]} \left(\theta_i^{[t+1]} - \theta^{[t]}\right), \tag{16}$$

Subsequently, it updates the first and second moments

$$m^{[t]} = \beta_1 m^{[t-1]} + (1-\beta_1)\Delta^{[t]}, \quad v^{[t]} = \beta_2 v^{[t-1]} + (1-\beta_2)(\Delta^{[t]})^2, \tag{17}$$

where $\beta_1, \beta_2 \in [0, 1)$ are decay rates. The server then performs an adaptive global update:

$$\theta^{[t+1]} = \theta^{[t]} + \eta m^{[t]} \Big/ \left(\sqrt{v^{[t]}} + \varepsilon_s\right), \tag{18}$$

where $\eta > 0$ is the server-side learning rate, and $\varepsilon_s > 0$ is a small constant for numerical stability.

3.3 Convergence Analysis

We first establish the necessary assumptions that have been widely adopted in the convergence analysis of stochastic optimization (Kingma & Ba, 2015; McMahan et al., 2017; Reddi et al., 2021; Jin et al., 2022; Khodadadian et al., 2022; Lan et al., 2025). We conduct our analysis on a convex region $\Theta \subseteq \mathbb{R}^d$ containing all global and local iterates, following the convention in stochastic optimization analyses (Berahas et al., 2024; Fang et al., 2024).

Assumption 3.1 (Lipschitz token log-policy)**.** For any input prompt q and corresponding response o, the log-probability $\log \pi_\theta(o_\ell \mid q, o_{<\ell})$ is finite and differentiable with respect to θ on Θ. Its gradient is Lipschitz continuous: there exists a constant $0 \leq L_1 < \infty$, such that

$$\|\nabla_\theta \log \pi_\theta(o_\ell \mid q, o_{<\ell}) - \nabla_{\theta'} \log \pi_{\theta'}(o_\ell \mid q, o_{<\ell})\|_2 \leq L_1 \|\theta - \theta'\|_2, \quad \forall \theta, \theta' \in \Theta \tag{19}$$

Assumption 3.2 (Bounded token policy score)**.** There exist $0 \leq S_2, S_\infty < \infty$ such that, for every relevant $(q, o_{<\ell})$, o_ℓ, and $\theta \in \Theta$,

$$\|\nabla_\theta \log \pi_\theta(o_\ell \mid q, o_{<\ell})\|_2 \leq S_2, \quad \|\nabla_\theta \log \pi_\theta(o_\ell \mid q, o_{<\ell})\|_\infty \leq S_\infty. \tag{20}$$

Assumption 3.3 (Bounded gradient variance)**.** There exists a constant $0 \leq \kappa_{\text{var}} < \infty$ such that

$$\sum_{i=1}^{N} p_i \|\nabla \mathcal{J}_i(\theta) - \nabla \mathcal{J}(\theta)\|_2^2 \leq \kappa_{\text{var}}^2, \quad \forall \theta \in \Theta. \tag{21}$$

where

$$\nabla \mathcal{J}_i(\theta) = \mathbb{E}_{q \sim \mathcal{D}_i,\, o \sim \pi_\theta(\cdot|q)}\Big[r(q,o) \sum_{\ell=1}^{|o|} \nabla_\theta \log \pi_\theta(o_\ell \mid q, o_{<\ell})\Big], \quad \nabla \mathcal{J}(\theta) = \sum_{i=1}^{N} p_i \nabla \mathcal{J}_i(\theta). \tag{22}$$

Assumption 3.4 (Bounded parameter drift from the reference policy)**.** The reference policy belongs to the same parameterized family as the optimized policy: $\pi_{\mathrm{ref}} = \pi_{\theta_{\mathrm{ref}}}$ for some $\theta_{\mathrm{ref}} \in \Theta$. There exists a finite constant $0 < L_{\mathrm{ref}} < \infty$, such that

$$\sup_{\theta \in \Theta} \|\theta - \theta_{\mathrm{ref}}\|_2 \leq L_{\mathrm{ref}}. \tag{23}$$

Based on the above assumptions, we establish the following convergence guarantee for FGRPO. Due to space limitations, we present only a simplified form of our main convergence result here; the formal theorem and its detailed proof are provided in Appendix A.

Theorem 3.1 (Convergence of FGRPO)**.** *Fix E, U, K, B, the initialization, and all problem and optimizer constants. Let $\overline{L} \triangleq \max\{L, 1\}$, and use the learning rates $\eta = \frac{1}{T^{1/4}}$ and $\hat{\eta} = \frac{1}{\overline{L}EUT^{1/4}}$. For Algorithm 1, when the communication horizon T is sufficiently long, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla \mathcal{J}(\theta^{[t]})\right\|_2^2\right] \leq \mathcal{O}(1/\sqrt{T}) + \mathcal{O}(1) \tag{24}$$

The above theorem characterize the finite-time convergence of FGRPO, separating an $\mathcal{O}(T^{-1/2})$ transient term from a residual upper bound that remains uniformly bounded as T increases. The transient term matches the standard rate for stochastic optimization (Shao et al., 2024; Pang & Jin, 2025). As detailed in Appendix A, the residual decomposition quantifies the contributions of client-gradient heterogeneity, GRPO update bias, and sampling noise, together with the effect of RPG-based weighting. Such residual terms also appear in bounds for state-of-the-art FedRL methods even with critics (Yang et al., 2024; Xie & Song, 2025), while establishing a residual that also decays with T may require stronger assumptions or sharper control of the underlying discrepancies.

4 Experiments

4.1 Experiment setup

We select **Qwen2.5-VL-3B-Instruct**, **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025b), **Qwen3-VL-4B-Instruct** (Bai et al., 2025a) and **Llama-3.2-11B-Vision-Instruct** (Grattafiori et al., 2024) to evaluate FGRPO across different scales, ranging from resource-efficient edge-level models to high-performance reasoning models. We evaluate FGRPO on the **OpenR1** (EvolvingLMMs-Lab, 2025) and **GEOQA** (Chen et al., 2025a) benchmarks. OpenR1 is a large-scale reasoning dataset with verified reasoning traces, including mathematical reasoning and other multi-step tasks, while GEOQA is a multimodal geometric question answering benchmark for numerical reasoning over both textual descriptions and visual diagrams. For OpenR1, we simulate data heterogeneity by partitioning samples into three difficulty tiers (*simple*, *medium*, and *hard*) according to reasoning trace length, and then allocating them to clients using a Dirichlet distribution $\mathsf{Dir}(\mu)$, where smaller μ induces more severe non-IID partitions. For GEOQA, we instead model domain heterogeneity by partitioning samples according to geometric primitives (*points*, *lines*, *circles*, and *polygons*) under the same Dirichlet-based strategy, such that different clients specialize in distinct geometric concepts. We set $\mu = 0.05$ for both datasets to create highly heterogeneous client distributions. Each dataset is split into 80% for training and 20% for testing.

We compare FGRPO with three representative FL baselines: **FedAvg** (McMahan et al., 2017), **FedProx** (Li et al., 2020), and **SCAFFOLD** (Karimireddy et al., 2020). Although these methods were not originally designed for GRPO, their mechanisms for handling non-IID data can be naturally adapted to our FGRPO framework. We evaluate the reasoning process by comparing the predicted answer against the ground-truth reference. A binary reward of 1 is assigned if the prediction matches the reference, and 0 otherwise. Consequently, we define the accuracy of a fine-tuned model as the aggregate proportion of these correct predictions across the test set. We implement these different algorithms on a cluster equipped with ten NVIDIA RTX Pro 6000 GPUs. To improve computational and communication efficiency, we incorporate Low-Rank Adaptation (LoRA) (Hu et al., 2022) into FGRPO. More details about the experiment settings can be found in the appendix (see Appendix B).

Table 1: Test accuracy (%) of different methods on the Open-R1 and GEOQA benchmarks. Results are reported as mean±std. **Bold** values indicate the best performance among decentralized methods, while underlined values denote the second-best results.

Setup		Open-R1				GEOQA				
Model	**Method**	**Simple**	**Medium**	**Hard**	**Total**	**Points**	**Lines**	**Circles**	**Polygons**	**Total**
Qwen2.5-3B	FedAvg	47.63±1.15	40.24±1.45	28.08±3.75	38.64±1.09	41.43±4.94	36.89±2.69	40.93±2.74	31.78±0.93	35.59±1.31
	FedProx	41.80±1.33	42.04±0.67	24.55±1.60	36.12±0.86	47.62±6.52	40.99±3.17	44.89±1.15	35.09±0.70	39.28±1.14
	SCAFFOLD	46.37±2.43	43.00±4.09	27.19±0.98	38.84±1.68	**51.90±3.53**	41.37±1.68	40.79±1.06	34.98±0.66	38.08±0.85
	FGRPO	**48.11±0.96**	**44.14±3.87**	**33.35±2.19**	**41.86±1.21**	47.62±4.12	**45.34±2.64**	**46.65±1.38**	**35.88±1.40**	**40.76±0.74**
Qwen2.5-7B	FedAvg	**60.90±0.98**	43.72±0.89	32.75±2.38	45.78±0.70	58.57±6.43	50.06±2.35	55.33±2.06	44.20±1.20	48.87±1.32
	FedProx	58.56±1.96	46.25±1.20	32.75±1.83	45.84±1.19	56.67±2.61	51.68±3.21	52.47±1.58	**45.84±1.68**	48.95±1.59
	SCAFFOLD	55.98±2.20	48.17±1.48	32.69±3.16	45.60±1.10	57.62±7.22	52.17±3.26	51.94±1.48	44.99±1.89	48.42±1.75
	FGRPO	55.44±1.68	**50.87±1.43**	**36.77±2.20**	**47.68±0.79**	**60.00±4.58**	**55.40±1.55**	**59.16±1.29**	45.34±2.17	**51.36±0.90**
Qwen3-4B	FedAvg	62.82±3.49	**49.37±2.38**	8.38±3.16	40.16±0.69	53.33±7.06	45.34±2.37	53.79±1.57	**51.15±2.12**	51.40±1.86
	FedProx	**65.83±1.61**	44.98±1.92	14.37±3.47	41.70±0.55	56.19±4.33	52.17±1.46	57.31±0.78	50.28±1.64	52.93±0.83
	SCAFFOLD	63.06±1.64	45.05±0.74	9.21±1.88	39.08±0.67	**60.48±2.71**	52.42±2.90	54.67±1.85	49.75±1.36	51.95±1.27
	FGRPO	65.47±1.18	45.29±2.51	**18.80±3.61**	**43.16±1.50**	53.81±3.98	**55.40±0.81**	**58.59±1.94**	50.83±1.28	**53.93±1.37**
Llama-3.2-11B	FedAvg	**49.25±2.03**	40.84±1.52	31.80±1.43	40.62±0.46	33.81±5.43	30.06±5.28	27.49±2.17	25.34±1.17	26.82±1.33
	FedProx	48.83±1.25	40.24±2.12	32.04±2.03	40.36±0.72	32.86±5.43	30.68±3.12	27.49±0.98	24.49±0.75	26.41±0.66
	SCAFFOLD	48.17±1.17	**42.04±2.16**	33.29±2.23	41.16±1.64	35.24±4.26	29.81±2.85	29.38±0.96	24.04±0.84	26.75±0.80
	FGRPO	49.07±1.87	41.26±1.51	**35.33±3.10**	**41.88±0.89**	**36.19±1.06**	**31.18±3.02**	**30.13±1.38**	**26.37±1.64**	**28.42±1.00**

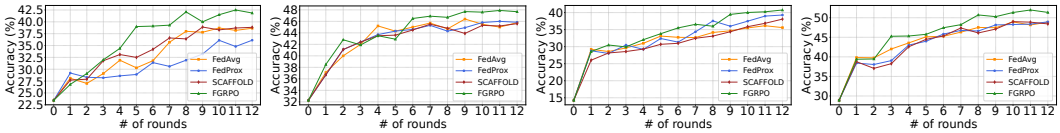
4.2 Experiment Results

Due to space limitations, we report results only for a five-client FL system with $\mu = 0.05$, $\lambda_{\text{base}} = 0.8$, $\tau_{\min} = 1.5$, $\tau_{\max} = 2.5$, $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.2$, and $\lambda_{\text{anneal}} = 0.1$. Additional experimental results, including the effects of varying the number of clients and data heterogeneity levels, extensions to different GRPO variants with RPG ablations, hyperparameter analysis, and resource consumption, are provided in Appendix C.

Table 1 shows the model accuracy of the different algorithms. On Open-R1, FGRPO obtains the highest total accuracy across all four backbone models. For Qwen2.5-3B, FGRPO improves the total accuracy from the strongest baseline of 38.84% to 41.86%, yielding a gain of 3.02%. For Qwen2.5-7B, FGRPO achieves 47.68%, outperforming the best baseline FedProx by 1.84%. The advantage remains consistent on larger and cross-family backbones. On Qwen3-4B, FGRPO reaches 43.16% total accuracy, surpassing the strongest baseline FedProx by 1.46%. This improvement is particularly pronounced on the challenging *hard* split, where FGRPO achieves 18.80%, substantially higher than FedProx at 14.37%. This indicates that the proposed RPG-based aggregation is especially beneficial for difficult reasoning examples, where clients may exhibit heterogeneous local optimization progress and reward distribution. On Llama-3.2-11B, FGRPO also achieves the best total accuracy of 41.88%, outperforming the strongest baseline SCAFFOLD by 0.72%. More importantly, FGRPO again shows clear advantages on the *hard* split, obtaining 35.33% compared with 33.29% from the best baseline. These results suggest that FGRPO is not only effective for Qwen-based models, but can also generalize to a distinct vision-language model family. The advantage of FGRPO is even more evident on the more challenging *hard* split of Open-R1. Compared with the strongest baseline, FGRPO improves the Hard accuracy by 5.27%, 4.02%, 4.43%, and 2.04% on Qwen2.5-3B, Qwen2.5-7B, Qwen3-4B, and Llama-3.2-11B, respectively.

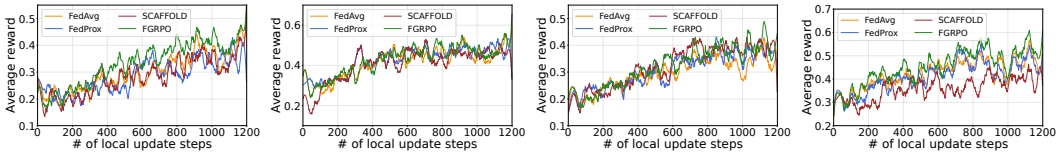
On GEOQA, FGRPO also consistently achieves the best total performance across all four models. For Qwen2.5-3B, FGRPO reaches 40.76% total accuracy, surpassing the best baseline FedProx by 1.48%. On Qwen2.5-7B, FGRPO further improves the total score to 51.36%, outperforming FedProx by 2.41%. The same trend holds on Qwen3-4B, where FGRPO obtains the highest total accuracy of 53.93%, improving over FedProx by 1.00%. Although FedAvg and SCAFFOLD perform competitively on some individual geometric categories, FGRPO achieves the best results on *lines* and *circles*, indicating stronger robustness on diverse geometric reasoning tasks. On Llama-3.2-11B, FGRPO achieves 28.42% total accuracy, surpassing FedAvg by 1.60%. Notably, FGRPO obtains the best performance across all GEOQA subcategories, including *points*, *lines*, *circles*, and *polygons*. This further confirms that the benefits of FGRPO are not limited to specific models or datasets, but remain effective across different backbone architectures and reasoning task structures.

The training trajectories in Fig. 1 further validate the effectiveness of FGRPO during federated optimization. For both Open-R1 and GEOQA, FGRPO generally reaches higher final accuracy and



(a) 3B model on OpenR1 (b) 7B model on OpenR1 (c) 3B model on GEOQA (d) 7B model on GEOQA

Figure 1: Convergence performance of different methods in terms of model accuracy with Qwen2.5-3B and Qwen2.5-7B models.



(a) 3B model on OpenR1 (b) 7B model on OpenR1 (c) 3B model on GEOQA (d) 7B model on GEOQA

Figure 2: Convergence performance of different algorithms in terms of average reward with Qwen2.5-3B and Qwen2.5-7B models.

maintains a stronger upward trend over communication rounds compared with the other baselines. In particular, its advantage becomes more pronounced in later rounds, suggesting that progress-aware aggregation supports sustained policy improvement under heterogeneous client data. The reward curves in Fig. 2 show a consistent pattern: FGRPO achieves more favorable reward trajectories across local update steps.

5 Brief Literature Survey

FedRL synergizes collaborative learning with sequential decision-making to enhance sample efficiency while preserving privacy. Recent theoretical advances have established rigorous convergence guarantees under data heterogeneity (Zhang et al., 2024) and Markovian sampling complexities (Khodadadian et al., 2022), with extensions to offline settings identifying sufficient conditions for global optimality (Woo et al., 2024). Beyond theory, research has addressed practical system constraints, including robustness against Byzantine failures (Fan et al., 2021; Fang et al., 2025) and asynchronous communication delays (Lan et al., 2025). Furthermore, managing heterogeneity remains a central focus, where methods utilizing momentum-based aggregation (Wang et al., 2024), policy distillation (Jiang et al., 2025), and shared representation learning (Xiong et al., 2025) have been proposed to stabilize updates across diverse environments.

GRPO eliminates the critic network to scale reasoning model fine-tuning via group-level reward normalization (Shao et al., 2024; Guo et al., 2025). To mitigate the high computational costs and lack of sensitivity in the original framework, recent works have introduced efficiency optimizations such as completion pruning (Lin et al., 2025) and semantic entropy integration (Chen et al., 2025b). Concurrently, precision-enhancing techniques like difficulty-aware advantage reweighting (Zhang & Zuo, 2025) and dynamic sampling systems (Yu et al., 2025) have been developed to stabilize long reasoning. Whereas these advancements implicitly assume centralized data availability, our proposed framework extends GRPO to distributed settings, enabling the collaborative training of reasoning LLMs across non-IID data. A comprehensive survey is provided in Appendix D.

6 Conclusion

In this paper, we introduced *federated group relative policy optimization* (FGRPO) for federated fine-tuning of reasoning-capable *large language models* (LLMs). FGRPO uses *relative performance gain* (RPG) to adapt aggregation across clients with heterogeneous data distributions. Our analysis bounds the average expected squared gradient norm by an $\mathcal{O}(T^{-1/2})$ transient term and a $\mathcal{O}(1)$ residual upper bound. On the Open-R1 and GEOQA benchmarks, FGRPO achieves the highest total accuracy among the evaluated federated baselines across all four backbone models. These results support the empirical value of RPG-based aggregation while clarifying the scope of the convergence guarantee.

AI USE STATEMENT

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE-Style Optimization for Learning from Human Feedback in LLMs. In *Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12248–12267, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Albert S. Berahas, Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient jacobians. *Mathematics of Operations Research*, 49(4):2212–2248, 2024. doi: 10.1287/moor.2021.0154.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than $3. `https://github.com/Deep-Agent/R1-V`, 2025a. Accessed: 2025-02-02.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. SEED-GRPO: Semantic Entropy Enhanced GRPO for Uncertainty-Aware Policy Optimization. *arXiv preprint arXiv:2505.12346*, 2025b.

EvolvingLMMs-Lab. Open-R1-Multimodal: A Fork to Add Multimodal Model Training to Open-R1. `https://github.com/EvolvingLMMs-Lab/open-r1-multimodal`, 2025.

Flint Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. Fault-Tolerant Federated Reinforcement Learning with Theoretical Guarantee. In *Proc. of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1007–1021, 2021.

Flint Xiaofeng Fan, Cheston Tan, Yew-Soon Ong, Roger Wattenhofer, and Wei Tsang Ooi. FedRLHF: A Convergence-Guaranteed Federated Framework for Privacy-Preserving and Personalized RLHF. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 713–721, 2025.

Minghong Fang, Xilong Wang, and Neil Zhenqiang Gong. Provably Robust Federated Reinforcement Learning. In *Proc. of the 2025 ACM on Web Conference (WWW)*, pp. 896–909, 2025.

Yuchen Fang, Sen Na, Michael W. Mahoney, and Mladen Kolar. Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *SIAM Journal on Optimization*, 34(2):2007–2037, 2024. doi: 10.1137/22M1537862.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature*, 645(8081):633–638, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. of the 10th International Conference on Learning Representations (ICLR)*, 2022.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 System Card. *arXiv preprint arXiv:2412.16720*, 2024.

Wenzheng Jiang, Ji Wang, Xiongtao Zhang, Weidong Bao, Cheston Tan, and Flint Xiaofeng Fan. FedHPD: Heterogeneous Federated Reinforcement Learning via Policy Distillation. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 2568–2570, 2025.

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated Reinforcement Learning with Environment Heterogeneity. In *Proc. of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 18–37, 2022.

Philip Jordan, Florian Grötschla, Flint Xiaofeng Fan, and Roger Wattenhofer. Decentralized Federated Policy Gradient with Byzantine Fault-Tolerance and Provably Fast Convergence. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 964–972, 2024.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proc. of the 37th International Conference on Machine Learning (ICML)*, pp. 5132–5143, 2020.

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling. In *Proc. of the 39th International Conference on Machine Learning (ICML)*, pp. 10997–11057, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proc. of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1008–1014, 1999.

Guangchen Lan, Dong-Jun Han, Abolfazl Hashemi, Vaneet Aggarwal, and Christopher Brinton. Asynchronous Federated Reinforcement Learning with Policy Gradient Updates: Algorithm Design and Convergence Analysis. In *Proc. of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. In *Proc. of the 3rd Conference on Machine Learning and Systems (MLSys)*, 2020.

Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. CPPO: Accelerating the Training of Group Relative Policy Optimization-Based Reasoning Models. *arXiv preprint arXiv:2503.22342*, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc. of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, 2017.

Lei Pang and Ruinan Jin. On the Theory and Practice of GRPO: A Trajectory-Corrected Approach with Fast Convergence. *arXiv preprint arXiv:2508.02833*, 2025.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive Federated Optimization. In *Proc. of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Desik Rengarajan, Nitin Ragothaman, Dileep Kalathil, and Srinivas Shakkottai. Federated Ensemble-Directed Offline Reinforcement Learning. In *Proc. of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

Han Wang, Sihong He, Zhili Zhang, Fei Miao, and James Anderson. Momentum for the Win: Collaborative Federated Reinforcement Learning across Heterogeneous Environments. In *Proc. of the 41st International Conference on Machine Learning (ICML)*, 2024.

Jiin Woo, Gauri Joshi, and Yuejie Chi. The Blessing of Heterogeneity in Federated Q-Learning: Linear Speedup and Beyond. In *Proc. of the 40th International Conference on Machine Learning (ICML)*, pp. 37157–37216, 2023.

Jiin Woo, Laixi Shi, Gauri Joshi, and Yuejie Chi. Federated Offline Reinforcement Learning: Collaborative Single-Policy Coverage Suffices. In *Proc. of the 41st International Conference on Machine Learning (ICML)*, 2024.

Zhijie Xie and Shenghui Song. The Actor-Critic Update Order Matters for PPO in Federated Reinforcement Learning. *arXiv preprint arXiv:2506.01261*, 2025.

Guojun Xiong, Shufan Wang, Daniel Jiang, and Jian Li. On the Linear Speedup of Personalized Federated Reinforcement Learning with Shared Representations. In *Proc. of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Tong Yang, Shicong Cen, Yuting Wei, Yuxin Chen, and Yuejie Chi. Federated Natural Policy Gradient and Actor Critic Methods for Multi-task Reinforcement Learning. In *Proc. of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

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.

Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-Time Analysis of On-Policy Heterogeneous Federated Reinforcement Learning. In *Proc. of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Jixiao Zhang and Chunsheng Zuo. GRPO-LEAD: A Difficulty-Aware Reinforcement Learning Approach for Concise Mathematical Reasoning in Language Models. *arXiv preprint arXiv:2504.09696*, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A Survey of Reinforcement Learning for Large Reasoning Models. *arXiv preprint arXiv:2509.08827*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group Sequence Policy Optimization. *arXiv preprint arXiv:2507.18071*, 2025.

APPENDIX CONTENTS

A Proof of Theorem 3.1 **14**

A.1 Auxiliary Lemmas 14

A.2 Main Proof 26

B More Information about Experiment Settings **32**

B.1 Implementation 32

B.2 Datasets and Baselines 33

C Supplementary Experiment Results **34**

C.1 Additional Comparison Results 34

C.2 Impact of the Number of Clients 34

C.3 Impact of Data Heterogeneity Levels 36

C.4 Extensions to GRPO-based Variants with RPG Ablations 37

C.5 Hyperparameter Sensitivity Analysis 38

C.6 Resource Consumption 39

D Comprehensive Literature Survey **40**

D.1 Federated Reinforcement Learning 40

D.2 GRPO 41

APPENDIX

A PROOF OF **THEOREM** 3.1

This appendix presents the proof of Theorem 3.1. In Sec. A.1, we establish smoothness and update bounds and decompose the effective update direction into the population gradient and errors due to local drift, RPG reweighting, GRPO surrogate bias, and sampling noise. In Sec. A.2, we incorporate server momentum and adaptive scaling into a one-round ascent bound and telescope it over communication rounds. Applying the prescribed step sizes then yields an $\mathcal{O}(T^{-1/2})$ transient term and a $\mathcal{O}(1)$ residual upper bound.

A.1 AUXILIARY LEMMAS

A.1.1 CONSEQUENCES OF THE ASSUMPTIONS

A bounded token policy score limits how much a token's log-probability can change as the parameters move away from the reference. Together with the bounded parameter drift, this yields a uniform bound on the ratio between reference and optimized token probabilities, as formalized below.

Lemma A.1 (Reference-policy coverage)**.** *Under Assumptions 3.1, 3.2, and 3.4, the reference policy and every policy π_θ, $\theta \in \Theta$, have common support on every relevant sampled token. Moreover,*

$$\exp(-S_2 L_{\mathrm{ref}}) \le \frac{\pi_{\mathrm{ref}}(o_\ell \mid q, o_{<\ell})}{\pi_\theta(o_\ell \mid q, o_{<\ell})} \le \exp(S_2 L_{\mathrm{ref}}), \qquad \forall \theta \in \Theta. \tag{25}$$

Proof. Fix a relevant context–token triple $x = (q, o_{<\ell}, o_\ell)$ and an arbitrary $\theta \in \Theta$, and define

$$f_x(\vartheta) \triangleq \log \pi_\vartheta(o_\ell \mid q, o_{<\ell}).$$

Assumption 3.1 states that f_x is a finite-valued differentiable function on Θ. Consequently, $\pi_\vartheta(o_\ell \mid q, o_{<\ell}) > 0$ for any $\vartheta \in \Theta$. In particular, the token has positive probability under both π_θ and $\pi_{\mathrm{ref}} = \pi_{\theta_{\mathrm{ref}}}$, which establishes common support.

It remains to bound the likelihood ratio. For $\lambda \in [0, 1]$, set

$$\tilde{\theta}(\lambda) \triangleq \theta_{\mathrm{ref}} + \lambda(\theta - \theta_{\mathrm{ref}}).$$

Since $\theta_{\mathrm{ref}},\ \theta \in \Theta$ and Θ is convex, $\tilde{\theta}(\lambda) \in \Theta$ for every $\lambda \in [0, 1]$. $f_x(\tilde{\theta}(\lambda))$ is differentiable, and the chain rule yields

$$\frac{d}{d\lambda} f_x(\tilde{\theta}(\lambda)) = \left\langle \nabla f_x(\tilde{\theta}(\lambda)), \theta - \theta_{\mathrm{ref}} \right\rangle. \tag{26}$$

We then have

$$\log \frac{\pi_\theta(o_\ell \mid q, o_{<\ell})}{\pi_{\mathrm{ref}}(o_\ell \mid q, o_{<\ell})} = f_x(\theta) - f_x(\theta_{\mathrm{ref}}) = \int_0^1 \left\langle \nabla f_x(\tilde{\theta}(\lambda)), \theta - \theta_{\mathrm{ref}} \right\rangle d\lambda. \tag{27}$$

Taking absolute values, applying Cauchy–Schwarz, and then using the bounds (20) from Assumption 3.2 and (23) from Assumption 3.4, respectively, yields

$$\begin{aligned} \left| \log \frac{\pi_{\mathrm{ref}}(o_\ell \mid q, o_{<\ell})}{\pi_\theta(o_\ell \mid q, o_{<\ell})} \right| &= \left| \int_0^1 \left\langle \nabla f_x(\tilde{\theta}(\lambda)), \theta - \theta_{\mathrm{ref}} \right\rangle d\lambda \right| \\ &\le \int_0^1 \|\nabla f_x(\tilde{\theta}(\lambda))\|_2 \|\theta - \theta_{\mathrm{ref}}\|_2 \, d\lambda \\ &\le S_2 \|\theta - \theta_{\mathrm{ref}}\|_2 \le S_2 L_{\mathrm{ref}}. \end{aligned} \tag{28}$$

Equivalently,

$$-S_2 L_{\mathrm{ref}} \le \log \frac{\pi_{\mathrm{ref}}(o_\ell \mid q, o_{<\ell})}{\pi_\theta(o_\ell \mid q, o_{<\ell})} \le S_2 L_{\mathrm{ref}}. \tag{29}$$

The exponential function is increasing, so exponentiating all three terms completes the proof. □

Define

$$r^*_{\max} \triangleq \max\{|r_{\min}|, |r_{\max}|\}, \qquad A_{\max} \triangleq \frac{r_{\max} - r_{\min}}{\varepsilon_A}. \tag{30}$$

For $u = 0, \ldots, U-2$, assume

$$R_0 = 0, \qquad R_{u+1} = R_u + \hat{\eta} S_2 \left(A_{\max} \exp(S_2 R_u) + \beta \exp(S_2 L_{\text{ref}})\right), \tag{31}$$

where recursion is empty when $U = 1$. Set

$$\begin{aligned} \rho_{\max} &\triangleq \exp(S_2 R_{U-1}), \\ G_2 &\triangleq S_2(A_{\max}\rho_{\max} + \beta \exp(S_2 L_{\text{ref}})), \\ G_\infty &\triangleq S_\infty(A_{\max}\rho_{\max} + \beta \exp(S_2 L_{\text{ref}})). \end{aligned} \tag{32}$$

The following lemma shows that the gradients of the local and global objectives change at most linearly with the parameters, while the local gradient norms are uniformly bounded and the global objective is bounded above. These properties allow us to control the objective change in each communication round and bound the cumulative improvement in the convergence proof.

Lemma A.2 (Smoothness and boundedness of the global and local objectives)**.** *Under Assumptions 3.1 and 3.2, together with the uniform reward and response-length bounds, for every $\theta, \theta' \in \Theta$,*

$$\|\nabla \mathcal{J}_i(\theta) - \nabla \mathcal{J}_i(\theta')\|_2 \le L\|\theta - \theta'\|_2, \qquad i = 1, \ldots, N, \tag{33}$$

$$\|\nabla \mathcal{J}(\theta) - \nabla \mathcal{J}(\theta')\|_2 \le L\|\theta - \theta'\|_2, \tag{34}$$

where

$$L \triangleq r^*_{\max}\big(\ell^2_{\max} S_2^2 + \ell_{\max} L_1\big). \tag{35}$$

Moreover,

$$\|\nabla \mathcal{J}_i(\theta)\|_2 \le \ell_{\max} r^*_{\max} S_2, \qquad \mathcal{J}(\theta) \le J^* \triangleq \sup_{\vartheta \in \Theta} \mathcal{J}(\vartheta) \le r_{\max}. \tag{36}$$

Proof. For any $\theta, \theta' \in \Theta$ and relevant (q, o), since $\nabla_\theta \log \pi_\theta(o \mid q) = \sum_{\ell=1}^{|o|} \nabla_\theta \log \pi_\theta(o_\ell \mid q, o_{<\ell})$, we have

$$\|\nabla_\theta \log \pi_\theta(o \mid q)\|_2 \le |o| S_2 \le \ell_{\max} S_2, \tag{37}$$

and

$$\|\nabla_\theta \log \pi_\theta(o \mid q) - \nabla_{\theta'} \log \pi_{\theta'}(o \mid q)\|_2 \le |o| L_1 \|\theta - \theta'\|_2 \le \ell_{\max} L_1 \|\theta - \theta'\|_2, \tag{38}$$

according to Assumptions 3.1 and 3.2.

Define $\tilde{\theta}(\lambda) = \theta' + \lambda(\theta - \theta')$ for $\lambda \in [0, 1]$. The convexity of Θ ensures $\tilde{\theta}(\lambda) \in \Theta$. Notice that $\tilde{\theta}(0) = \theta'$ and $\tilde{\theta}(1) = \theta$, and $d\tilde{\theta}(\lambda)/d\lambda = \theta - \theta'$. For every relevant response o, the token log-probabilities $\log \pi_\theta(o_\ell \mid q, o_{<\ell})$, and thus $\log \pi_\theta(o \mid q)$, are finite and differentiable on Θ. Therefore, $\nabla_\theta \pi_\theta(o \mid q) = \pi_\theta(o \mid q) \nabla_\theta \log \pi_\theta(o \mid q)$. Applying the chain rule gives

$$\frac{d}{d\lambda} \pi_{\tilde{\theta}(\lambda)}(o \mid q) = \left\langle \nabla_{\tilde{\theta}(\lambda)} \pi_{\tilde{\theta}(\lambda)}(o \mid q), \frac{d\tilde{\theta}(\lambda)}{d\lambda} \right\rangle = \pi_{\tilde{\theta}(\lambda)}(o \mid q) \left\langle \nabla_{\tilde{\theta}(\lambda)} \log \pi_{\tilde{\theta}(\lambda)}(o \mid q), \theta - \theta' \right\rangle. \tag{39}$$

By applying (37), we have

$$\begin{aligned} \sum_o |\pi_\theta(o \mid q) - \pi_{\theta'}(o \mid q)| &= \sum_o \left| \int_0^1 \frac{d}{d\lambda} \pi_{\tilde{\theta}(\lambda)}(o \mid q)\, d\lambda \right| \\ &\le \sum_o \int_0^1 \left| \frac{d}{d\lambda} \pi_{\tilde{\theta}(\lambda)}(o \mid q) \right| d\lambda \\ &= \int_0^1 \sum_o \pi_{\tilde{\theta}(\lambda)}(o \mid q) \left| \left\langle \nabla_{\tilde{\theta}(\lambda)} \log \pi_{\tilde{\theta}(\lambda)}(o \mid q), \theta - \theta' \right\rangle \right| d\lambda \\ &\le \int_0^1 \sum_o \pi_{\tilde{\theta}(\lambda)}(o \mid q) \ell_{\max} S_2 \|\theta - \theta'\|_2 \, d\lambda \\ &= \ell_{\max} S_2 \|\theta - \theta'\|_2. \end{aligned} \tag{40}$$

Expanding the expectation over responses in the definition of the local objective $\mathcal{J}_i(\theta)$ gives

$$\mathcal{J}_i(\theta) = \mathbb{E}_{q\sim\mathcal{D}_i}\left[\sum_o \pi_\theta(o \mid q) r(q,o)\right].$$

The response space is finite for each prompt because the vocabulary and response-length cap are finite. Moreover, $\sum_o |r(q,o)| \|\nabla_\theta \pi_\theta(o \mid q)\|_2 \le r^*_{\max}\ell_{\max}S_2$ uniformly over q and $\theta \in \Theta$. This bound justifies differentiation under the prompt expectation. Since $r(q,o)$ is independent of θ,

$$\nabla\mathcal{J}_i(\theta) = \mathbb{E}_{q\sim\mathcal{D}_i}\left[\sum_o r(q,o)\nabla_\theta\pi_\theta(o \mid q)\right] = \mathbb{E}_{q\sim\mathcal{D}_i}\left[\sum_o r(q,o)\pi_\theta(o \mid q)\nabla_\theta \log \pi_\theta(o \mid q)\right] \tag{41}$$

where the second equality uses $\nabla_\theta\pi_\theta(o \mid q) = \pi_\theta(o \mid q)\nabla_\theta \log \pi_\theta(o \mid q)$. By adding and subtracting $\pi_{\theta'}(o \mid q)\nabla_\theta \log \pi_\theta(o \mid q)$ inside the difference of the two gradients $\nabla\mathcal{J}_i(\theta)$ and $\nabla\mathcal{J}_i(\theta')$, we have

$$\begin{aligned}
&\|\nabla\mathcal{J}_i(\theta) - \nabla\mathcal{J}_i(\theta')\|_2 \\
=&\left\|\mathbb{E}_{q\sim\mathcal{D}_i}\left[\sum_o r(q,o)\big(\pi_\theta(o \mid q)\nabla_\theta \log \pi_\theta(o \mid q) - \pi_{\theta'}(o \mid q)\nabla_{\theta'} \log \pi_{\theta'}(o \mid q)\big)\right]\right\|_2 \\
\le&\mathbb{E}_{q\sim\mathcal{D}_i}\left[\sum_o |r(q,o)| \cdot |\pi_\theta(o \mid q) - \pi_{\theta'}(o \mid q)| \cdot \|\nabla_\theta \log \pi_\theta(o \mid q)\|_2\right] \\
&+ \mathbb{E}_{q\sim\mathcal{D}_i}\left[\sum_o |r(q,o)| \cdot \pi_{\theta'}(o \mid q) \cdot \|\nabla_\theta \log \pi_\theta(o \mid q) - \nabla_{\theta'} \log \pi_{\theta'}(o \mid q)\|_2\right] \\
\le&r^*_{\max}\ell_{\max}S_2\big(\ell_{\max}S_2\|\theta-\theta'\|_2\big) + r^*_{\max}\ell_{\max}L_1\|\theta-\theta'\|_2 \\
=&r^*_{\max}\big(\ell^2_{\max}S_2^2 + \ell_{\max}L_1\big)\|\theta-\theta'\|_2 = L\|\theta-\theta'\|_2.
\end{aligned} \tag{42}$$

where the first inequality uses Jensen's inequality and the triangle inequality after the stated addition and subtraction, while the second inequality uses $|r(q,o)| \le r^*_{\max}$, $\sum_o \pi_\theta(o \mid q) = 1$, Eqs. (37), (40), and (38). This proves (33).

Since $\nabla\mathcal{J} = \sum_i p_i \nabla\mathcal{J}_i$ and $\sum_i p_i = 1$,

$$\|\nabla\mathcal{J}(\theta) - \nabla\mathcal{J}(\theta')\|_2 \le \sum_i p_i L\|\theta-\theta'\|_2 = L\|\theta-\theta'\|_2,$$

which proves (34). Finally, the score-function identity and Eq. (37) imply

$$\|\nabla\mathcal{J}_i(\theta)\|_2 \le \mathbb{E}[|r(q,o)|\|\nabla_\theta \log \pi_\theta(o \mid q)\|_2] \le \ell_{\max}r^*_{\max}S_2.$$

Moreover, bounded rewards give $\mathcal{J}(\vartheta) \le r_{\max}$ for every $\vartheta \in \Theta$; taking the supremum completes Eq. (36). □

In each rollout cycle (t,e), the local updates are initialized at the frozen rollout-cycle parameter $\theta_i^{[t,e]}$. The following lemma shows that the reuse updates remain in a bounded neighborhood of this frozen parameter, and that the resulting GRPO directions are bounded.

Lemma A.3 (Uniform bound for the sampled GRPO directions)**.** *Under Assumptions 3.1– 3.4, for every client i, round t, rollout cycle e, reuse step $u \in \{0,\dots,U-1\}$, and relevant $(q, o_{<\ell}, o_\ell)$, the following bounds hold:*

$$\left\|\theta_i^{[t,e,u]} - \theta_i^{[t,e]}\right\|_2 \le R_u, \tag{43}$$

$$0 < \frac{\pi_{\theta_i^{[t,e,u]}}(o_\ell \mid q, o_{<\ell})}{\pi_{\theta_i^{[t,e]}}(o_\ell \mid q, o_{<\ell})} \le \rho_{\max}, \tag{44}$$

$$\left\|g_i^{[t,e,u]}\right\|_2 \le G_2, \ \left\|g_i^{[t,e,u]}\right\|_\infty \le G_\infty. \tag{45}$$

Here R_u, $\rho_{\max}$, G_2, and G_∞ are the derived quantities in Eqs. (31)–(32)*.*

Proof. Fixing (i,t,e), we have $R_0 = 0$ when $u = 0$ at the initialization. In the following, we first derive a recurrence for $\|\theta_i^{[t,e,u]} - \theta_i^{[t,e]}\|_2$ directly from the local updates, and then compare it with the deterministic sequence R_u in (31). For any (i,t,e,u) and $(q, o_{<\ell}, o_\ell)$, define

$$\tilde{\theta}(\lambda) \triangleq \theta_i^{[t,e]} + \lambda\big(\theta_i^{[t,e,u]} - \theta_i^{[t,e]}\big), \qquad \lambda \in [0,1].$$

Since Θ is convex, $\tilde{\theta}(\lambda) \in \Theta$ for every $\lambda \in [0,1]$. Assumption 3.1 therefore makes $\log \pi_{\tilde{\theta}(\lambda)}(o_\ell \mid q, o_{<\ell})$ continuous on $[0,1]$ and differentiable on $(0,1)$. By the chain rule,

$$\frac{d}{d\lambda} \log \pi_{\tilde{\theta}(\lambda)}(o_\ell \mid q, o_{<\ell}) = \left\langle \nabla_{\tilde{\theta}(\lambda)} \log \pi_{\tilde{\theta}(\lambda)}(o_\ell \mid q, o_{<\ell}),\ \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\rangle.$$

The mean-value theorem gives some $\lambda^* \in (0,1)$, such that

$$\begin{aligned}
\left| \log \frac{\pi_{\theta_i^{[t,e,u]}}(o_\ell \mid q, o_{<\ell})}{\pi_{\theta_i^{[t,e]}}(o_\ell \mid q, o_{<\ell})} \right| &= \left| \left\langle \nabla_{\tilde{\theta}(\lambda^*)} \log \pi_{\tilde{\theta}(\lambda^*)}(o_\ell \mid q, o_{<\ell}), \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\rangle \right| \\
&\le \left\| \nabla_{\tilde{\theta}(\lambda^*)} \log \pi_{\tilde{\theta}(\lambda^*)}(o_\ell \mid q, o_{<\ell}) \right\|_2 \cdot \left\| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\|_2 \\
&\le S_2 \left\| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\|_2, \qquad (46)
\end{aligned}$$

where the first inequality is Cauchy–Schwarz, while the second one uses Assumption 3.2 at $\tilde{\theta}(\lambda^*) \in \Theta$. Therefore, we have

$$\frac{\pi_{\theta_i^{[t,e,u]}}(o_\ell \mid q, o_{<\ell})}{\pi_{\theta_i^{[t,e]}}(o_\ell \mid q, o_{<\ell})} \le \exp\left(S_2 \|\theta_i^{[t,e,u]} - \theta_i^{[t,e]}\|_2 \right). \qquad (47)$$

According to Eq. (30), we have

$$\left| A_{i,k}^{[t,e]}(q) \right| \le A_{\max}. \qquad (48)$$

For a fixed sampled token, abbreviate $A = A_{i,k}^{[t,e]}(q)$, $\Phi(\theta) = \Phi_{\theta,\theta_i^{[t,e]}}(o_{k,\ell} \mid q, o_{k,<\ell})$ and $\rho(\theta) = \rho_{\theta,\theta_i^{[t,e]}}(o_{k,\ell} \mid q, o_{k,<\ell})$. The clipped advantage term is

$$\Phi(\theta) = \begin{cases} A\rho(\theta), & A > 0 \text{ and } \rho(\theta) < 1 + \varepsilon_{\text{clip}}, \\ A(1+\varepsilon_{\text{clip}}), & A > 0 \text{ and } \rho(\theta) \ge 1 + \varepsilon_{\text{clip}}, \\ A\rho(\theta), & A < 0 \text{ and } \rho(\theta) > 1 - \varepsilon_{\text{clip}}, \\ A(1-\varepsilon_{\text{clip}}), & A < 0 \text{ and } \rho(\theta) \le 1 - \varepsilon_{\text{clip}}, \\ 0, & A = 0. \end{cases} \qquad (49)$$

The nonconstant branch has gradient $A\rho(\theta)\nabla_\theta \log \pi_\theta(o_{k,\ell} \mid q, o_{k,<\ell})$, and the constant branch has gradient zero. Therefore, the selected direction is

$$\nabla_\theta \Phi(\theta) = \mathbb{I}_A(\rho(\theta)) A \rho(\theta) \nabla_\theta \log \pi_\theta(o_{k,\ell} \mid q, o_{k,<\ell}). \qquad (50)$$

where

$$\mathbb{I}_A\left(\rho(\theta)\right) \triangleq \begin{cases} 1, & A > 0 \text{ and } \rho(\theta) < 1 + \varepsilon_{\text{clip}}, \\ 1, & A < 0 \text{ and } \rho(\theta) > 1 - \varepsilon_{\text{clip}}, \\ 0, & \text{otherwise.} \end{cases} \qquad (51)$$

At a switching threshold the coefficient is zero, so this formula specifies the direction even when the ordinary gradient of Φ does not exist. Consequently,

$$\begin{aligned}
&\left\| \nabla_\theta \Phi_{\theta,\theta_i^{[t,e]}}(o_\ell \mid q, o_{<\ell}) \right\|_2 \\
\le &\left| A_{i,k}^{[t,e]}(q) \right| \cdot \rho_{\theta,\theta_i^{[t,e]}}(o_\ell \mid q, o_{<\ell}) \cdot \|\nabla_\theta \log \pi_\theta(o_\ell \mid q, o_{<\ell})\|_2 \\
\le &A_{\max} S_2 \rho_{\theta,\theta_i^{[t,e]}}(o_\ell \mid q, o_{<\ell}),
\end{aligned}$$

where the second inequality uses the score bound (20) and advantage bound (48). At $\theta = \theta_i^{[t,e,u]}$, by substituting (47) into the above inequality, we have

$$\left\| \nabla_\theta \Phi_{\theta,\theta_i^{[t,e]}}(o_\ell \mid q, o_{<\ell}) \Big|_{\theta=\theta_i^{[t,e,u]}} \right\|_2 \le A_{\max} S_2 \exp\left(S_2 \left\| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\|_2 \right). \tag{52}$$

According to Eq. (5), differentiating the KL term $\widehat{\mathbb{D}}_{k,\ell}$ gives

$$\nabla_\theta \widehat{\mathbb{D}}_{k,\ell}(\theta; q) = \left(1 - \frac{\pi_{\text{ref}}(o_\ell \mid q, o_{<\ell})}{\pi_\theta(o_\ell \mid q, o_{<\ell})} \right) \nabla_\theta \log \pi_\theta(o_\ell \mid q, o_{<\ell}).$$

According to Lemma A.1 and $\exp(S_2 L_{\text{ref}}) \ge 1$, we have

$$\left| 1 - \frac{\pi_{\text{ref}}(o_\ell \mid q, o_{<\ell})}{\pi_\theta(o_\ell \mid q, o_{<\ell})} \right| \le \exp(S_2 L_{\text{ref}})$$

Combining this inequality with the score bound (20) gives

$$\left\| \nabla_\theta \widehat{\mathbb{D}}_{k,\ell}(\theta; q) \right\|_2 \le S_2 \exp(S_2 L_{\text{ref}}). \tag{53}$$

The local direction is an average over sampled tokens, responses, and prompt groups. Applying the triangle inequality to this average gives

$$\begin{aligned}
&\left\| g_i^{[t,e,u]} \right\|_2 \\
&\le \frac{1}{BK} \sum_{q \in \mathcal{B}_i^{[t,e]}} \sum_{k=1}^{K} \frac{1}{|o_k|} \sum_{\ell=1}^{|o_k|} \left(\left\| \nabla_\theta \Phi_{\theta,\theta_i^{[t,e]}}(o_{k,\ell} \mid q, o_{k,<\ell}) \right\|_2 + \beta \left\| \nabla_\theta \widehat{\mathbb{D}}_{k,\ell}(\theta; q) \right\|_2 \right) \Big|_{\theta=\theta_i^{[t,e,u]}} \\
&\le S_2 \left(A_{\max} \exp\left(S_2 \left\| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\|_2 \right) + \beta \exp(S_2 L_{\text{ref}}) \right).
\end{aligned} \tag{54}$$

The final bound follows from (52) and (53), since the averaging weights sum to one. For $u \le U - 2$, the update rule (11) then gives

$$\begin{aligned}
&\left\| \theta_i^{[t,e,u+1]} - \theta_i^{[t,e]} \right\|_2 \\
&= \left\| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} + \hat{\eta} g_i^{[t,e,u]} \right\|_2 \\
&\le \left\| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\|_2 + \hat{\eta} \left\| g_i^{[t,e,u]} \right\|_2 \\
&\le \left\| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\|_2 + \hat{\eta} S_2 \left(A_{\max} \exp\left(S_2 \| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \|_2 \right) + \beta \exp(S_2 L_{\text{ref}}) \right).
\end{aligned} \tag{55}$$

The displacement is zero at $u = 0$. Since the right-hand side is nondecreasing in the preceding displacement, induction using (31) proves (43). When $U = 1$, only this initial case is needed.

Substituting Eq. (43) into Eq. (46) now yields

$$\left| \log \frac{\pi_{\theta_i^{[t,e,u]}}(o_\ell \mid q, o_{<\ell})}{\pi_{\theta_i^{[t,e]}}(o_\ell \mid q, o_{<\ell})} \right| \le S_2 \left\| \theta_i^{[t,e,u]} - \theta_i^{[t,e]} \right\|_2 \le S_2 R_u. \tag{56}$$

Since R_u is nondecreasing such that $R_u \le R_{U-1}$, exponentiation gives Eq. (44). Similarly, Eq. (54) gives

$$\left\| g_i^{[t,e,u]} \right\|_2 \le S_2 \left(A_{\max} \exp(S_2 R_u) + \beta \exp(S_2 L_{\text{ref}}) \right) \le G_2, \tag{57}$$

using Eq. (32).

Finally, applying the same estimates to the clipped advantage and KL regularization terms in the infinity norm, using $\| \nabla_\theta \log \pi_\theta(o_\ell \mid q, o_{<\ell}) \|_\infty \le S_\infty$, yields

$$\left\| g_i^{[t,e,u]} \right\|_\infty \le S_\infty \left(A_{\max} \rho_{\max} + \beta \exp(S_2 L_{\text{ref}}) \right) = G_\infty,$$

completing the proof. □

A.1.2 AN ORACLE FOR PPO-STYLE SAMPLE REUSE

Let $\mathcal{F}_{i,t,e}$ denote client i's history before sampling block (t,e), including the broadcast model and all preceding local blocks. For fixed i,t, these histories form an increasing filtration; $\mathcal{F}_{i,t,e+1}$ includes the samples and all reuse iterates from block (t,e), also when $e=E-1$. Conditional on this history, Algorithm 1 samples B independent prompt groups, each comprising $q\sim\mathcal{D}_i$ and K conditionally independent responses from the frozen policy $\pi_{\theta_i^{[t,e]}}(\cdot\mid q)$. Successive blocks are conditionally fresh; clients need not be independent, and group-normalized advantages may remain dependent.

Once the history and the sampled block are fixed, the update rule (12) determines every iterate within the block, with the boundary convention (51) specifying the update direction at clipping thresholds. Because these updates use measurable operations, each iterate is a measurable function of the history and the sampled block. Define the averages and centered residual in (t,e) by

$$\bar{g}_i^{[t,e]} \triangleq \frac{1}{U}\sum_{u=0}^{U-1} g_i^{[t,e,u]}, \tag{58}$$

$$\nabla\overline{\mathcal{J}}_i^{[t,e]} \triangleq \frac{1}{U}\sum_{u=0}^{U-1} \nabla\mathcal{J}_i(\theta_i^{[t,e,u]}), \tag{59}$$

$$\xi_i^{[t,e]} \triangleq \bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]} - \mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]} \;\middle|\; \mathcal{F}_{i,t,e}\right]. \tag{60}$$

The direction $\bar{g}_i^{[t,e]}$ decomposes into this comparator, its conditional bias, and a conditionally centered residual:

$$\bar{g}_i^{[t,e]} = \nabla\overline{\mathcal{J}}_i^{[t,e]} + \mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]} \;\middle|\; \mathcal{F}_{i,t,e}\right] + \xi_i^{[t,e]}, \quad \mathbb{E}\left[\xi_i^{[t,e]} \;\middle|\; \mathcal{F}_{i,t,e}\right] = 0. \tag{61}$$

For a run of T rounds, define deterministic uniform bounds on the conditional bias magnitude and conditional variance by

$$\delta_G \triangleq \sup_{i,t,e} \operatorname{ess\,sup} \left\| \mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]} \;\middle|\; \mathcal{F}_{i,t,e}\right] \right\|_2, \tag{62}$$

$$\nu_G^2 \triangleq \sup_{i,t,e} \operatorname{ess\,sup} \mathbb{E}\left[\left\| \xi_i^{[t,e]} \right\|_2^2 \;\middle|\; \mathcal{F}_{i,t,e}\right]. \tag{63}$$

Here the outer supremum ranges over $i=1,\dots,N$, $t=0,\dots,T-1$,and $e=0,\dots,E-1$. For each fixed (i,t,e), the essential supremum is taken over the randomness in the conditioning history $\mathcal{F}_{i,t,e}$. Thus, δ_G and ν_G^2 are deterministic constants that bound the corresponding conditional quantities almost surely, uniformly over clients and blocks.

Lemma A.4 (Block-oracle bias and variance)**.** *Under Assumptions 3.1– 3.4 and the sampling protocol stated above, the following general bounds hold for every $U\geq 1$:*

$$\delta_G \leq G_2 + \ell_{\max} r^*_{\max} S_2, \qquad \nu_G^2 \leq (G_2 + \ell_{\max} r^*_{\max} S_2)^2. \tag{64}$$

For the special case $U=1$, the bias bound remains valid, while the variance bound can be sharpened to

$$\nu_G^2 \leq G_2^2/B. \tag{65}$$

Proof. Fix a client i and rollout cycle (t,e). Conditional on $\mathcal{F}_{i,t,e}$, the starting parameter $\theta_i^{[t,e]}$ is fixed, whereas the prompts and responses sampled in this cycle remain random. Consequently, the update directions $g_i^{[t,e,u]}$ and the subsequent parameters $\theta_i^{[t,e,u]}$ for $u\geq 1$, which depend on these samples, are also random.

According to the definitions of $\bar{g}_i^{[t,e]}$ and $\overline{\mathcal{J}}_i^{[t,e]}$ in Eqs. (58) and (59), respectively, for every realization, we have

$$\begin{aligned}
\left\| \bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]} \right\|_2 &= \left\| \frac{1}{U}\sum_{u=0}^{U-1} \left(g_i^{[t,e,u]} - \nabla\mathcal{J}_i(\theta_i^{[t,e,u]}) \right) \right\|_2 \\
&\leq \frac{1}{U}\sum_{u=0}^{U-1} \left(\left\| g_i^{[t,e,u]} \right\|_2 + \left\| \nabla\mathcal{J}_i(\theta_i^{[t,e,u]}) \right\|_2 \right) \\
&\leq G_2 + \ell_{\max} r^*_{\max} S_2.
\end{aligned} \tag{66}$$

where the second inequality uses (36) in Lemma A.2 and (45) in Lemma A.3. Conditional Jensen's inequality and (66) then give

$$\left\|\mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\middle|\mathcal{F}_{i,t,e}\right]\right\|_2 \le \mathbb{E}\left[\left\|\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\right\|_2\middle|\mathcal{F}_{i,t,e}\right] \le G_2 + \ell_{\max} r^*_{\max} S_2.$$

According to the definition of $\xi_i^{[t,e]}$ in (60), we have

$$\begin{aligned}\left\|\xi_i^{[t,e]}\right\|_2^2 = &\left\|\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\right\|_2^2 + \left\|\mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\middle|\mathcal{F}_{i,t,e}\right]\right\|_2^2 \\ &- 2\left\langle \bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}, \mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\middle|\mathcal{F}_{i,t,e}\right]\right\rangle. \end{aligned} \tag{67}$$

$\mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\middle|\mathcal{F}_{i,t,e}\right]$ is $\mathcal{F}_{i,t,e}$-measurable, and is thus fixed when conditioning on the history. Therefore, the third term in (67) satisfies

$$\mathbb{E}\Big[\Big\langle \bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}, \mathbb{E}\Big[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\Big|\mathcal{F}_{i,t,e}\Big]\Big\rangle \;\Big|\; \mathcal{F}_{i,t,e}\Big] = \left\|\mathbb{E}\Big[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\Big|\mathcal{F}_{i,t,e}\Big]\right\|_2^2. \tag{68}$$

Its squared norm is also unchanged by conditioning on $\mathcal{F}_{i,t,e}$. Taking conditional expectations of the expansion gives

$$\begin{aligned}\mathbb{E}\Big[\left\|\xi_i^{[t,e]}\right\|_2^2\Big|\mathcal{F}_{i,t,e}\Big] =&\mathbb{E}\Big[\|\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\|_2^2 \mid \mathcal{F}_{i,t,e}\Big] - \left\|\mathbb{E}\Big[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\Big|\mathcal{F}_{i,t,e}\Big]\right\|_2^2 \\ \le&\mathbb{E}\Big[\|\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]}\|_2^2\Big|\mathcal{F}_{i,t,e}\Big] \\ \le&(G_2 + \ell_{\max} r^*_{\max} S_2)^2. \end{aligned} \tag{69}$$

where the first inequality drops a nonpositive squared-mean term, and the last uses (66). For any i and (t,e), the preceding estimates show that the conditional-bias norm is at most $G_2 + \ell_{\max} r^*_{\max} S_2$ and the conditional second moment of $\xi_i^{[t,e]}$ is at most $(G_2 + \ell_{\max} r^*_{\max} S_2)^2$. These constants do not depend on the client, block, or realized history. They therefore also bound the essential suprema defining δ_G and ν_G^2 in Eqs. (62)–(63). Here an essential supremum is the smallest upper bound that holds with probability one; values on probability-zero events do not affect it. This proves (64).

For $U = 1$, the parameter at which the direction is evaluated is $\theta_i^{[t,e]}$, which is fixed conditional on $\mathcal{F}_{i,t,e}$. Index the sampled prompt groups by $b = 1, \ldots, B$. Write q_b for the b-th prompt and $o_{b,k}$ for its k-th independently generated response. Its contribution is

$$\begin{aligned}g_{i,b}^{[t,e,0]} \triangleq \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|o_{b,k}|}\sum_{\ell=1}^{|o_{b,k}|}&\left(A_{i,b,k}^{[t,e]} - \beta\left(1 - \frac{\pi_{\mathrm{ref}}(o_{b,k,\ell} \mid q_b, o_{b,k,<\ell})}{\pi_{\theta_i^{[t,e]}}(o_{b,k,\ell} \mid q_b, o_{b,k,<\ell})}\right)\right) \\ &\nabla_\theta \log \pi_\theta(o_{b,k,\ell} \mid q_b, o_{b,k,<\ell})|_{\theta=\theta_i^{[t,e]}}, \end{aligned} \tag{70}$$

where $A_{i,b,k}^{[t,e]}$ is the advantage computed within prompt group b. The behavior ratio equals one at this first update, so the clipping indicator contributes $A_{i,b,k}^{[t,e]}$ for every sign of the advantage. Hence

$$g_i^{[t,e,0]} = \frac{1}{B}\sum_{b=1}^{B} g_{i,b}^{[t,e,0]}. \tag{71}$$

The B contributions are conditionally independent because each uses only its own prompt and responses. Furthermore, according to Eqs. (48), (25) in Lemma A.1, and (20) in Assumption 3.2, we have

$$\left\|g_{i,b}^{[t,e,0]}\right\|_2 \le S_2(A_{\max} + \beta\exp(S_2 L_{\mathrm{ref}})) = G_2,$$

where $\rho_{\max} = 1$ for $U = 1$.

The population comparator is $\nabla\overline{\mathcal{J}}_i^{[t,e]} = \nabla\mathcal{J}_i(\theta_i^{[t,e]})$ and is $\mathcal{F}_{i,t,e}$-measurable. It therefore cancels from the centered residual. Defining $\widetilde{g}_{i,b}^{[t,e,0]} = g_{i,b}^{[t,e,0]} - \mathbb{E}\left[g_{i,b}^{[t,e,0]}\middle|\mathcal{F}_{i,t,e}\right]$ gives $\xi_i^{[t,e]} = B^{-1}\sum_b \widetilde{g}_{i,b}^{[t,e,0]}$ and

$$\mathbb{E}\left[\left\|\widetilde{g}_{i,b}^{[t,e,0]}\right\|_2^2\middle|\mathcal{F}_{i,t,e}\right] = \mathbb{E}\left[\left\|g_{i,b}^{[t,e,0]}\right\|_2^2\middle|\mathcal{F}_{i,t,e}\right] - \left\|\mathbb{E}\left[g_{i,b}^{[t,e,0]}\middle|\mathcal{F}_{i,t,e}\right]\right\|_2^2 \le G_2^2.$$

For distinct indices $b \neq b'$, conditional independence and zero conditional means make the cross term vanish. This remains true if the two sampled prompt values happen to coincide. Consequently,

$$\mathbb{E}\left[\left\|\xi_i^{[t,e]}\right\|_2^2 \middle| \mathcal{F}_{i,t,e}\right] = \frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left[\left\|\widetilde{g}_{i,b}^{[t,e,0]}\right\|_2^2 \middle| \mathcal{F}_{i,t,e}\right] \leq \frac{G_2^2}{B}. \tag{72}$$

Taking the essential supremum over histories and the supremum over clients and blocks proves (65). □

A.1.3 Policy-gradient heterogeneity and adaptive aggregation

RPG aggregation favors clients with larger relative performance gains, so its weights may differ from the objective weights p_i. Bounded rewards and positive volatility and temperature floors limit how concentrated these weights can become. The following lemma bounds both the amplification of each client's weight relative to p_i and the overall weighted squared deviation from the objective weights.

Lemma A.5 (RPG density-ratio bound)**.** *For Algorithm 1, under Assumptions 3.1–3.4, for every round t and every client i,*

$$\left|h_i^{[t]}\right| \leq \frac{r_{\max}-r_{\min}}{\sigma_{\min}+\varepsilon_{\mathrm{rpg}}}, \qquad \frac{\omega_i^{[t]}}{p_i} \leq C_{\mathrm{RPG}}, \qquad \sum_{i=1}^{N}\frac{(\omega_i^{[t]}-p_i)^2}{p_i} \leq C_{\mathrm{RPG}}-1. \tag{73}$$

where

$$C_{\mathrm{RPG}} \triangleq \frac{1}{Np_{\min}}\exp\left(\frac{2(r_{\max}-r_{\min})}{(\sigma_{\min}+\varepsilon_{\mathrm{rpg}})\tau_{\min}}\right). \tag{74}$$

Proof. Since both $\bar{r}_i^{[t]}$ and $\varphi_i^{[t-1]}$ lie in $[r_{\min}, r_{\max}]$, the first result follows from $\sigma_i^{[t]} \geq \sigma_{\min}$, according to the definition of $h_i^{[t]}$ in Eq. (14). Next, since $\tau^{[t]} \geq \tau_{\min} > 0$, for each client i, we have

$$\left|\frac{h_i^{[t]}}{\tau^{[t]}}\right| \leq \frac{r_{\max}-r_{\min}}{(\sigma_{\min}+\varepsilon_{\mathrm{rpg}})\tau_{\min}}.$$

Hence, according to the definition of $\omega_i^{[t]}$ in Eq. (15), we have

$$\omega_i^{[t]} = \frac{\exp(h_i^{[t]}/\tau^{[t]})}{\sum_{j=1}^{N}\exp(h_j^{[t]}/\tau^{[t]})} \leq \frac{1}{N}\exp\left(\frac{2(r_{\max}-r_{\min})}{(\sigma_{\min}+\varepsilon_{\mathrm{rpg}})\tau_{\min}}\right).$$

Since $p_i \geq p_{\min} > 0$, division by p_i yields

$$\frac{\omega_i^{[t]}}{p_i} \leq \frac{1}{Np_{\min}}\exp\left(\frac{2(r_{\max}-r_{\min})}{(\sigma_{\min}+\varepsilon_{\mathrm{rpg}})\tau_{\min}}\right) = C_{\mathrm{RPG}}.$$

This proves the second result.

Finally,

$$\sum_i \frac{(\omega_i^{[t]}-p_i)^2}{p_i} = \sum_i \frac{(\omega_i^{[t]})^2}{p_i} - 1 \leq \left(\max_i \frac{\omega_i^{[t]}}{p_i}\right)\sum_i \omega_i^{[t]} - 1 \leq C_{\mathrm{RPG}} - 1. \tag{75}$$

□

Bounded local update directions limit how far each client can move in a round. Since aggregation and the server moment recursions use weighted averages, the aggregated displacement and both moments remain bounded. Together with the positive stabilizer ε_s, this ensures that the preconditioner scales each coordinate by a factor that is uniformly bounded above and away from zero.

Lemma A.6 (Bounds for the server moments and preconditioner)**.** *For Algorithm 1, under Assumptions 3.1–3.4, for every round t,*

$$\left\|\theta_i^{[t+1]}-\theta^{[t]}\right\|_2 \leq \hat{\eta}EUG_2, \quad \left\|\Delta^{[t]}\right\|_2 \leq \hat{\eta}EUG_2, \quad \left\|m^{[t]}\right\|_2 \leq \hat{\eta}EUG_2, \tag{76}$$

$$\left\|\theta_i^{[t+1]}-\theta^{[t]}\right\|_\infty \leq \hat{\eta}EUG_\infty, \quad 0 \leq v_j^{[t]} \leq \hat{\eta}^2E^2U^2G_\infty^2 \quad (j=1,\ldots,d). \tag{77}$$

Consequently, for

$$P^{[t]} \triangleq \operatorname{diag}\left(\frac{1}{\sqrt{v^{[t]}}+\varepsilon_s}\right), \tag{78}$$

one has

$$\frac{1}{\hat{\eta}EUG_\infty+\varepsilon_s} I \preceq P^{[t]} \preceq \frac{1}{\varepsilon_s} I. \tag{79}$$

Proof. Each client starts round t at $\theta_i^{[t,0,0]} = \theta^{[t]}$ and performs EU local updates. Summing the recursion (11) over these updates gives

$$\theta_i^{[t+1]} - \theta^{[t]} = \hat{\eta} \sum_{e=0}^{E-1} \sum_{u=0}^{U-1} g_i^{[t,e,u]}.$$

According to Lemma A.3, for either $\ell = 2$ or $\ell = \infty$, we have

$$\left\|\theta_i^{[t+1]} - \theta^{[t]}\right\|_\ell \le \hat{\eta} \sum_{e=0}^{E-1} \sum_{u=0}^{U-1} \left\|g_i^{[t,e,u]}\right\|_\ell \le \hat{\eta}EUG_\ell.$$

Recall that $\Delta^{[t]}$ denotes the weighted average of these client displacements, as defined in (16). Since the RPG weights satisfy $\omega_i^{[t]} \ge 0$ and $\sum_i \omega_i^{[t]} = 1$, for both norms we have

$$\left\|\Delta^{[t]}\right\|_\ell = \left\|\sum_{i=1}^N \omega_i^{[t]} (\theta_i^{[t+1]} - \theta^{[t]})\right\|_\ell \le \sum_{i=1}^N \omega_i^{[t]} \left\|\theta_i^{[t+1]} - \theta^{[t]}\right\|_\ell \le \hat{\eta}EUG_\ell.$$

In particular, $\left|\Delta_j^{[t]}\right| \le \hat{\eta}EUG_\infty$ for every coordinate j. These bounds hold for every round and every realization of the samples and RPG weights.

We next prove the moment bounds by induction. They hold at initialization because $m^{[-1]} = v^{[-1]} = 0$. Suppose

$$\left\|m^{[t-1]}\right\|_2 \le \hat{\eta}EUG_2, \qquad 0 \le v_j^{[t-1]} \le \hat{\eta}^2E^2U^2G_\infty^2 \quad (j = 1, \dots, d).$$

The first-moment recursion in (17), together with $0 \le \beta_1 < 1$, gives

$$\left\|m^{[t]}\right\|_2 \le \beta_1 \left\|m^{[t-1]}\right\|_2 + (1-\beta_1) \left\|\Delta^{[t]}\right\|_2 \le \beta_1\hat{\eta}EUG_2 + (1-\beta_1)\hat{\eta}EUG_2 = \hat{\eta}EUG_2.$$

Similarly, the second-moment recursion and $0 \le \beta_2 < 1$ imply

$$0 \le v_j^{[t]} = \beta_2 v_j^{[t-1]} + (1-\beta_2)(\Delta_j^{[t]})^2 \le \beta_2\hat{\eta}^2E^2U^2G_\infty^2 + (1-\beta_2)\hat{\eta}^2E^2U^2G_\infty^2 = \hat{\eta}^2E^2U^2G_\infty^2.$$

This completes the induction and proves Eqs. (76)–(77).

Finally, $0 \le v_j^{[t]} \le \hat{\eta}^2E^2U^2G_\infty^2$ gives $0 \le \sqrt{v_j^{[t]}} \le \hat{\eta}EUG_\infty$. Since $\varepsilon_s > 0$, each diagonal entry of the preconditioner satisfies

$$\frac{1}{\hat{\eta}EUG_\infty+\varepsilon_s} \le P_{jj}^{[t]} = \frac{1}{\sqrt{v_j^{[t]}}+\varepsilon_s} \le \frac{1}{\varepsilon_s}.$$

Because $P^{[t]}$ is diagonal, its eigenvalues are precisely these diagonal entries. Thus $(\hat{\eta}EUG_\infty + \varepsilon_s)^{-1} I \preceq P^{[t]} \preceq \varepsilon_s^{-1} I$, proving Eq. (79). □

A.1.4 The effective round direction

Each client starts round t at $\theta_i^{[t,0,0]} = \theta^{[t]}$ and performs E rollout cycles, each containing U local updates. Summing the local update recursion (11) over these EU updates telescopes the parameter increments:

$$\theta_i^{[t+1]} - \theta^{[t]} = \hat{\eta} \sum_{e=0}^{E-1} \sum_{u=0}^{U-1} g_i^{[t,e,u]} = \hat{\eta}U \sum_{e=0}^{E-1} \bar{g}_i^{[t,e]},$$

where the last equality uses $\bar{g}_i^{[t,e]} = U^{-1}\sum_{u=0}^{U-1} g_i^{[t,e,u]}$ defined in (58). Substituting this displacement into the server aggregation rule (16) and dividing by the common scale $\hat{\eta}EU$ gives the normalized round direction:

$$\overline{g}^{[t]} \triangleq \frac{\Delta^{[t]}}{\hat{\eta}EU} = \sum_{i=1}^{N} \omega_i^{[t]} \frac{1}{E} \sum_{e=0}^{E-1} \bar{g}_i^{[t,e]}. \tag{80}$$

Introduce the following four residuals:

$$\gamma_{\mathrm{drift}}^{[t]} \triangleq \sum_i \omega_i^{[t]} \frac{1}{EU} \sum_{e=0}^{E-1}\sum_{u=0}^{U-1} \big(\nabla \mathcal{J}_i(\theta_i^{[t,e,u]}) - \nabla \mathcal{J}_i(\theta^{[t]})\big), \tag{81}$$

$$\gamma_{\mathrm{RPG}}^{[t]} \triangleq \sum_i (\omega_i^{[t]} - p_i) \nabla \mathcal{J}_i(\theta^{[t]}), \tag{82}$$

$$\gamma_{\mathrm{bias}}^{[t]} \triangleq \sum_i \omega_i^{[t]} \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla \overline{\mathcal{J}}_i^{[t,e]} \;\middle|\; \mathcal{F}_{i,t,e}\right], \tag{83}$$

$$\gamma_{\mathrm{noise}}^{[t]} \triangleq \sum_i \omega_i^{[t]} \frac{1}{E} \sum_{e=0}^{E-1} \xi_i^{[t,e]}. \tag{84}$$

The drift residual $\gamma_{\mathrm{drift}}^{[t]}$ captures the change in population gradients as local iterates move away from the round's initial global model $\theta^{[t]}$. The RPG residual $\gamma_{\mathrm{RPG}}^{[t]}$ captures the change in the aggregated gradient caused by replacing the objective weights p_i with the adaptive weights $\omega_i^{[t]}$. The bias residual $\gamma_{\mathrm{bias}}^{[t]}$ aggregates the conditional mean discrepancies between the block GRPO directions and the averaged population gradients along the local update trajectories. The noise residual $\gamma_{\mathrm{noise}}^{[t]}$ aggregates the remaining sampling fluctuations of these block discrepancies after their conditional means have been subtracted.

Lemma A.7 (Round-direction decomposition)**.** *For Algorithm 1, under Assumptions 3.1–3.4 and the sampling protocol of Section A.1.2, the normalized round direction admits the following decomposition:*

$$\overline{g}^{[t]} = \nabla \mathcal{J}(\theta^{[t]}) + \gamma^{[t]}, \qquad \gamma^{[t]} \triangleq \gamma_{\mathrm{drift}}^{[t]} + \gamma_{\mathrm{RPG}}^{[t]} + \gamma_{\mathrm{bias}}^{[t]} + \gamma_{\mathrm{noise}}^{[t]}. \tag{85}$$

Moreover, the following mean-square bounds hold, where expectations are over all prompt and response sampling through round t, including the resulting randomness in the iterates and RPG weights:

$$\mathbb{E}\left[\left\|\gamma_{\mathrm{drift}}^{[t]}\right\|_2^2\right] \le \frac{L^2\hat{\eta}^2E^2U^2G_2^2}{3}, \tag{86}$$

$$\mathbb{E}\left[\left\|\gamma_{\mathrm{RPG}}^{[t]}\right\|_2^2\right] \le (C_{\mathrm{RPG}} - 1)\kappa_{\mathrm{var}}^2, \tag{87}$$

$$\mathbb{E}\left[\left\|\gamma_{\mathrm{bias}}^{[t]}\right\|_2^2\right] \le \delta_G^2 \le (G_2 + \ell_{\max} r_{\max}^* S_2)^2, \tag{88}$$

$$\mathbb{E}\left[\left\|\gamma_{\mathrm{noise}}^{[t]}\right\|_2^2\right] \le \frac{C_{\mathrm{RPG}}\nu_G^2}{E} \le \frac{C_{\mathrm{RPG}}}{E}(G_2 + \ell_{\max} r_{\max}^* S_2)^2. \tag{89}$$

Finally,

$$\mathbb{E}\left[\left\|\gamma^{[t]}\right\|_2^2\right] \le 4C_{\mathrm{FGRPO}}. \tag{90}$$

where

$$C_{\mathrm{FGRPO}} \triangleq \frac{L^2\hat{\eta}^2E^2U^2G_2^2}{3} + (C_{\mathrm{RPG}} - 1)\kappa_{\mathrm{var}}^2 + \delta_G^2 + \frac{C_{\mathrm{RPG}}\nu_G^2}{E}. \tag{91}$$

Proof. Unrolling the local updates gives

$$\theta_i^{[t+1]} - \theta^{[t]} = \hat{\eta} \sum_{e=0}^{E-1}\sum_{u=0}^{U-1} g_i^{[t,e,u]} = \hat{\eta}EU \frac{1}{E} \sum_{e=0}^{E-1} \bar{g}_i^{[t,e]}. \tag{92}$$

Substituting Eq. (61) into Eq. (80) and collecting the conditional bias and noise terms gives

$$\overline{g}^{[t]} = \sum_i \omega_i^{[t]} \frac{1}{E} \sum_{e=0}^{E-1} \nabla \overline{\mathcal{J}}_i^{[t,e]} + \gamma_{\text{bias}}^{[t]} + \gamma_{\text{noise}}^{[t]}.$$

To separate local drift from client reweighting, we expand the block comparator using (59) such that

$$\begin{aligned}
\sum_i \omega_i^{[t]} \frac{1}{E} \sum_{e=0}^{E-1} \nabla \overline{\mathcal{J}}_i^{[t,e]} &= \sum_i \omega_i^{[t]} \frac{1}{EU} \sum_{e=0}^{E-1} \sum_{u=0}^{U-1} \nabla \mathcal{J}_i(\theta_i^{[t,e,u]}) \\
&= \gamma_{\text{drift}}^{[t]} + \sum_i \omega_i^{[t]} \nabla \mathcal{J}_i(\theta^{[t]}) \\
&= \gamma_{\text{drift}}^{[t]} + \sum_i p_i \nabla \mathcal{J}_i(\theta^{[t]}) + \sum_i (\omega_i^{[t]} - p_i) \nabla \mathcal{J}_i(\theta^{[t]}) \\
&= \gamma_{\text{drift}}^{[t]} + \nabla \mathcal{J}(\theta^{[t]}) + \gamma_{\text{RPG}}^{[t]}.
\end{aligned}$$

where the second equality uses the definition of $\gamma_{\text{drift}}^{[t]}$; the third splits each aggregation weight as $\omega_i^{[t]} = p_i + (\omega_i^{[t]} - p_i)$; the last equality follows from $\nabla \mathcal{J}(\theta^{[t]}) = \sum_i p_i \nabla \mathcal{J}_i(\theta^{[t]})$ and the definition of $\gamma_{\text{RPG}}^{[t]}$. Combining the two decompositions yields (85).

At the start of round t, client i has $\theta_i^{[t,0,0]} = \theta^{[t]}$. Before step (e, u), it has completed $eU + u$ local updates. Each update has norm at most $\hat{\eta} G_2$ by Lemma A.3. Therefore,

$$\left\| \theta_i^{[t,e,u]} - \theta^{[t]} \right\|_2 \leq \hat{\eta} \left(\sum_{e'=0}^{e-1} \sum_{u'=0}^{U-1} \left\| g_i^{[t,e',u']} \right\|_2 + \sum_{u'=0}^{u-1} \left\| g_i^{[t,e,u']} \right\|_2 \right) \leq \hat{\eta}(eU + u) G_2. \tag{93}$$

Applying Jensen's inequality to (81), then taking expectations, gives

$$\begin{aligned}
\mathbb{E}\left[\left\| \gamma_{\text{drift}}^{[t]} \right\|_2^2 \right] &\leq \mathbb{E}\left[\sum_i \frac{\omega_i^{[t]}}{EU} \sum_{e=0}^{E-1} \sum_{u=0}^{U-1} \left\| \nabla \mathcal{J}_i(\theta_i^{[t,e,u]}) - \nabla \mathcal{J}_i(\theta^{[t]}) \right\|_2^2 \right] \\
&\leq L^2 \mathbb{E}\left[\sum_i \frac{\omega_i^{[t]}}{EU} \sum_{e=0}^{E-1} \sum_{u=0}^{U-1} \left\| \theta_i^{[t,e,u]} - \theta^{[t]} \right\|_2^2 \right] \\
&\leq L^2 \hat{\eta}^2 G_2^2 \mathbb{E}\left[\sum_i \frac{\omega_i^{[t]}}{EU} \sum_{e=0}^{E-1} \sum_{u=0}^{U-1} (eU + u)^2 \right] \\
&= \frac{L^2 \hat{\eta}^2 G_2^2}{EU} \sum_{k=0}^{EU-1} k^2 \\
&\leq \frac{L^2 \hat{\eta}^2 E^2 U^2 G_2^2}{3},
\end{aligned} \tag{94}$$

Here, the second inequality uses L-smoothness of each $\mathcal{J}_i$, the third uses the bound (93), the equality reindexes the local substeps by $k = eU + u \in \{0, \ldots, EU - 1\}$ and uses $\sum_i \omega_i^{[t]} = 1$, and the last inequality uses

$$\frac{1}{EU} \sum_{k=0}^{EU-1} k^2 = \frac{(EU-1)(2EU-1)}{6} \leq \frac{E^2 U^2}{3}.$$

The same estimates hold for each realized trajectory before taking expectations, so they also establish the drift bound.

Since $\sum_i \omega_i^{[t]} = \sum_i p_i = 1$, we have $\sum_i (\omega_i^{[t]} - p_i) = 0$. Subtracting the global gradient from each client gradient therefore leaves the RPG residual unchanged. Since $p_i > 0$ for every client, we can write

$$\begin{aligned}
\gamma_{\text{RPG}}^{[t]} &= \sum_i (\omega_i^{[t]} - p_i) \big(\nabla \mathcal{J}_i(\theta^{[t]}) - \nabla \mathcal{J}(\theta^{[t]}) \big) \\
&= \sum_i \frac{\omega_i^{[t]} - p_i}{\sqrt{p_i}} \left(\sqrt{p_i} \big(\nabla \mathcal{J}_i(\theta^{[t]}) - \nabla \mathcal{J}(\theta^{[t]}) \big) \right).
\end{aligned}$$

By applying Cauchy–Schwarz and then taking expectations, we have

$$\mathbb{E}\left[\left\|\gamma_{\mathrm{RPG}}^{[t]}\right\|_2^2\right] \le \mathbb{E}\left[\left(\sum_i \frac{(\omega_i^{[t]}-p_i)^2}{p_i}\right)\left(\sum_i p_i \left\|\nabla\mathcal{J}_i(\theta^{[t]})-\nabla\mathcal{J}(\theta^{[t]})\right\|_2^2\right)\right] \le (C_{\mathrm{RPG}}-1)\kappa_{\mathrm{var}}^2.$$

The first factor is at most $C_{\mathrm{RPG}}-1$ by Lemma A.5; the second is at most κ_{var}^2 by Assumption 3.3. The bounds hold before taking expectations and therefore imply the stated mean-square bound.

For the bias residual, define

$$b_i^{[t,e]} \triangleq \mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]} \;\middle|\; \mathcal{F}_{i,t,e}\right].$$

By (62), $\|b_i^{[t,e]}\|_2 \le \delta_G$ holds almost surely for every client and block. Since the weights $\omega_i^{[t]}/E$ sum to one over i, e, Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\left[\left\|\gamma_{\mathrm{bias}}^{[t]}\right\|_2^2\right] &= \mathbb{E}\left[\left\|\sum_i \frac{\omega_i^{[t]}}{E}\sum_{e=0}^{E-1} b_i^{[t,e]}\right\|_2^2\right] \\ &\le \mathbb{E}\left[\sum_i \frac{\omega_i^{[t]}}{E}\sum_{e=0}^{E-1}\left\|b_i^{[t,e]}\right\|_2^2\right] \\ &\le \delta_G^2\,\mathbb{E}\left[\sum_i \frac{\omega_i^{[t]}}{E}\sum_{e=0}^{E-1} 1\right] = \delta_G^2.\end{aligned}$$

The same calculation without the outer expectations gives $\|\gamma_{\mathrm{bias}}^{[t]}\|_2^2 \le \delta_G^2$. Substituting (64) proves (88).

For the noise residual, the RPG weights may depend on the same samples as the block noises. We first use Jensen's inequality and the bound $\omega_i^{[t]} \le C_{\mathrm{RPG}} p_i$ shown in (73) to obtain

$$\begin{aligned}\left\|\gamma_{\mathrm{noise}}^{[t]}\right\|_2^2 &= \left\|\sum_i \omega_i^{[t]}\frac{1}{E}\sum_{e=0}^{E-1}\xi_i^{[t,e]}\right\|_2^2 \\ &\le \sum_i \omega_i^{[t]}\left\|\frac{1}{E}\sum_{e=0}^{E-1}\xi_i^{[t,e]}\right\|_2^2 \\ &\le C_{\mathrm{RPG}}\sum_i p_i\left\|\frac{1}{E}\sum_{e=0}^{E-1}\xi_i^{[t,e]}\right\|_2^2. \end{aligned} \tag{95}$$

This replaces the random aggregation weights by the fixed objective weights before any cancellation of noise terms is used.

For each client, the definition (60) gives $\xi_i^{[t,e]} = \bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]} - b_i^{[t,e]}$. Since $b_i^{[t,e]}$ is $\mathcal{F}_{i,t,e}$-measurable,

$$\mathbb{E}\left[\xi_i^{[t,e]} \;\middle|\; \mathcal{F}_{i,t,e}\right] = \mathbb{E}\left[\bar{g}_i^{[t,e]} - \nabla\overline{\mathcal{J}}_i^{[t,e]} \;\middle|\; \mathcal{F}_{i,t,e}\right] - b_i^{[t,e]} = 0.$$

Moreover, taking an outer expectation of the conditional second-moment bound (63) yields

$$\mathbb{E}\left[\left\|\xi_i^{[t,e]}\right\|_2^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\xi_i^{[t,e]}\right\|_2^2 \;\middle|\; \mathcal{F}_{i,t,e}\right]\right] \le \mathbb{E}[\nu_G^2] = \nu_G^2. \tag{96}$$

Here the first equality is the tower property, and the last uses the fact that ν_G^2 is deterministic. In particular, the noises have finite second moments, so their pairwise inner products are integrable.

Once block e is completed, its samples and all reuse iterates belong to $\mathcal{F}_{i,t,e+1}$, so $\xi_i^{[t,e]}$ is $\mathcal{F}_{i,t,e+1}$-measurable. For $e < e'$, the increasing histories satisfy $\mathcal{F}_{i,t,e+1} \subseteq \mathcal{F}_{i,t,e'}$. The tower property and the rule for taking measurable factors outside a conditional expectation therefore give

$$\mathbb{E}\left[\left\langle \xi_i^{[t,e]}, \xi_i^{[t,e']}\right\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\left\langle \xi_i^{[t,e]}, \xi_i^{[t,e']}\right\rangle \;\middle|\; \mathcal{F}_{i,t,e'}\right]\right] = \mathbb{E}\left[\left\langle \xi_i^{[t,e]}, \mathbb{E}\left[\xi_i^{[t,e']} \;\middle|\; \mathcal{F}_{i,t,e'}\right]\right\rangle\right] = 0,$$

where the last equality uses the conditional centering of the later block noise. Thus, the cross terms vanish in expectation even though successive blocks need not be independent. Expanding the squared norm of the average block noise now gives

$$\mathbb{E}\left[\left\|\frac{1}{E}\sum_{e=0}^{E-1}\xi_i^{[t,e]}\right\|_2^2\right] = \frac{1}{E^2}\sum_{e=0}^{E-1}\mathbb{E}\left[\left\|\xi_i^{[t,e]}\right\|_2^2\right] + \frac{2}{E^2}\sum_{0\le e<e'\le E-1}\underbrace{\mathbb{E}\left[\left\langle \xi_i^{[t,e]}, \xi_i^{[t,e']}\right\rangle\right]}_{=0}$$
$$= \frac{1}{E^2}\sum_{e=0}^{E-1}\mathbb{E}\left[\left\|\xi_i^{[t,e]}\right\|_2^2\right] \le \frac{E\nu_G^2}{E^2} = \frac{\nu_G^2}{E}. \tag{97}$$

where the inequality uses (96).

Taking expectations in (95) and using $\sum_i p_i = 1$, we obtain

$$\mathbb{E}\left[\left\|\gamma_{\text{noise}}^{[t]}\right\|_2^2\right] \le C_{\text{RPG}}\sum_i p_i\,\mathbb{E}\left[\left\|\frac{1}{E}\sum_{e=0}^{E-1}\xi_i^{[t,e]}\right\|_2^2\right] \le C_{\text{RPG}}\sum_i p_i\frac{\nu_G^2}{E} = \frac{C_{\text{RPG}}\nu_G^2}{E}.$$

Substituting (64) proves the second inequality in (89). This argument requires neither independence between clients nor independence between the RPG weights and the block noises.

Finally, taking the expectation for $\gamma^{[t]}$ in (85) yields

$$\mathbb{E}\left[\|\gamma^{[t]}\|_2^2\right] \le 4\Big(\mathbb{E}\left[\|\gamma_{\text{drift}}^{[t]}\|_2^2\right] + \mathbb{E}\left[\|\gamma_{\text{RPG}}^{[t]}\|_2^2\right] + \mathbb{E}\left[\|\gamma_{\text{bias}}^{[t]}\|_2^2\right] + \mathbb{E}\left[\|\gamma_{\text{noise}}^{[t]}\|_2^2\right]\Big)$$
$$\le 4\left(\frac{L^2\hat{\eta}^2E^2U^2G_2^2}{3} + (C_{\text{RPG}}-1)\kappa_{\text{var}}^2 + \delta_G^2 + \frac{C_{\text{RPG}}\nu_G^2}{E}\right)$$
$$= 4C_{\text{FGRPO}},$$

where the last equality is the definition (91). This proves (90). □

A.2 Main Proof

We first present a full version of our main result in Theorem A.1, and then give a detaild proof.

Theorem A.1 (Finite-time stationarity bound for FGRPO)**.** *Fix E, U, K, B, the initialization, and all problem and optimizer constants except the step sizes. Let $\overline{L} \triangleq \max\{L, 1\}$ and choose a communication horizon satisfying*

$$T \ge T_0 \triangleq \max\left\{1, \left(\frac{2(U-1)S_2^2(A_{\max} + \beta\exp(S_2L_{\text{ref}}))}{\overline{L}EU}\right)^4\right\}. \tag{98}$$

Suppose the domain condition on Θ and Assumptions 3.1– 3.4 hold with the same constants for all such horizons, and use the sampling protocol of Section A.1.2. For Algorithm 1, use the constant-within-run step sizes

$$\eta = T^{-1/4}, \qquad \hat{\eta} = \frac{1}{\overline{L}EUT^{1/4}}. \tag{99}$$

Define the constants

$$G_2^* \triangleq S_2(2A_{\max} + \beta\exp(S_2L_{\text{ref}})), \quad G_\infty^* \triangleq S_\infty(2A_{\max} + \beta\exp(S_2L_{\text{ref}})). \tag{100}$$

Algorithm 1 satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{J}(\theta^{[t]})\|_2^2 \le \frac{C_{\text{tr}}}{\sqrt{T}} + C_{\text{res}}, \tag{101}$$

where the transient and residual constants are given by

$$C_{\text{res}} \triangleq \frac{8(\varepsilon_s + G_\infty^*/\overline{L})^2}{\varepsilon_s^2(1-\beta_1)}\left[(C_{\text{RPG}}+1)\kappa_{\text{var}}^2 + \left(1 + \frac{C_{\text{RPG}}}{E}\right)(G_2^* + \ell_{\max}r_{\max}^*S_2)^2\right],$$
$$C_{\text{tr}} \triangleq \frac{\varepsilon_s + G_\infty^*/\overline{L}}{1-\beta_1}\left(2\overline{L}(J^* - \mathcal{J}(\theta^{[0]})) + \frac{L(G_2^*)^2}{\varepsilon_s^2\overline{L}}\right)$$
$$+ \frac{8(\varepsilon_s + G_\infty^*/\overline{L})^2L^2(G_2^*)^2}{\varepsilon_s^2(1-\beta_1)\overline{L}^2}\left(\frac{1}{3} + \frac{\beta_1(1+\beta_1)}{4\varepsilon_s^2(1-\beta_1)^2}\right). \tag{102}$$

Proof. We first derive a bound for any finite horizon and arbitrary positive step sizes for which the stated assumptions hold. We then analyze the choice (99) for general $T \geq 1$ and finally impose $T \geq T_0$ to obtain the constants in the theorem. Repeatedly substituting the first-moment recursion $m^{[k]} = \beta_1 m^{[k-1]} + (1-\beta_1)\Delta^{[k]}$ from (17), for $k = t, t-1, \ldots, 0$, gives

$$m^{[t]} = \beta_1^{t+1} m^{[-1]} + (1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \Delta^{[k]}.$$

The initialization $m^{[-1]} = 0$ eliminates the first term. Substituting $\Delta^{[k]} = \hat{\eta} EU \overline{g}^{[k]}$ from (80) then yields

$$m^{[t]} = \hat{\eta} EU (1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \overline{g}^{[k]}. \tag{103}$$

Define

$$e^{[k,t]} \triangleq \nabla \mathcal{J}(\theta^{[k]}) - \nabla \mathcal{J}(\theta^{[t]}) + \gamma^{[k]}. \tag{104}$$

Lemma A.7 then yields

$$m^{[t]} = \hat{\eta} EU \left((1-\beta_1^{t+1}) \nabla \mathcal{J}(\theta^{[t]}) + (1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} e^{[k,t]} \right). \tag{105}$$

Lemma A.6 implies

$$\left\| \theta^{[j+1]} - \theta^{[j]} \right\|_2 \leq \frac{\eta \hat{\eta} EUG_2}{\varepsilon_s}. \tag{106}$$

Therefore, L-smoothness, the triangle inequality, and (90) give

$$\begin{aligned} \mathbb{E}\left[\left\| e^{[k,t]} \right\|_2^2 \right] &\leq 2L^2 \mathbb{E}\left[\left\| \theta^{[k]} - \theta^{[t]} \right\|_2^2 \right] + 2\mathbb{E}\left[\left\| \gamma^{[k]} \right\|_2^2 \right] \\ &\leq \frac{2L^2 \eta^2 \hat{\eta}^2 E^2 U^2 G_2^2}{\varepsilon_s^2} (t-k)^2 + 8C_{\mathrm{FGRPO}}. \end{aligned} \tag{107}$$

We next use the momentum decomposition to bound the alignment of the server update with the current gradient. Substituting (105) and applying the preconditioner bounds in (79) gives

$$\begin{aligned} &\frac{\langle \nabla \mathcal{J}(\theta^{[t]}), P^{[t]} m^{[t]} \rangle}{\hat{\eta} EU} \\ =&(1-\beta_1^{t+1}) \left\langle \nabla \mathcal{J}(\theta^{[t]}), P^{[t]} \nabla \mathcal{J}(\theta^{[t]}) \right\rangle + (1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \left\langle \nabla \mathcal{J}(\theta^{[t]}), P^{[t]} e^{[k,t]} \right\rangle \\ \geq& \frac{1-\beta_1^{t+1}}{\hat{\eta} EUG_\infty + \varepsilon_s} \left\| \nabla \mathcal{J}(\theta^{[t]}) \right\|_2^2 - \frac{\left\| \nabla \mathcal{J}(\theta^{[t]}) \right\|_2}{\varepsilon_s} (1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \left\| e^{[k,t]} \right\|_2. \end{aligned} \tag{108}$$

Here the lower preconditioner bound controls the positive quadratic term. For the error terms, we use Cauchy–Schwarz together with the upper bound $\|P^{[t]} e^{[k,t]}\|_2 \leq \|e^{[k,t]}\|_2 / \varepsilon_s$. Using Young's inequality to separate the gradient norm from the error sum, the second term in (108) can be bounded by

$$\begin{aligned} &\frac{\left\| \nabla \mathcal{J}(\theta^{[t]}) \right\|_2 (1-\beta_1)}{\varepsilon_s} \sum_{k=0}^{t} \beta_1^{t-k} \left\| e^{[k,t]} \right\|_2 \\ \leq& \frac{1-\beta_1^{t+1}}{2(\hat{\eta} EUG_\infty + \varepsilon_s)} \left\| \nabla \mathcal{J}(\theta^{[t]}) \right\|_2^2 + \frac{\hat{\eta} EUG_\infty + \varepsilon_s}{2\varepsilon_s^2 (1-\beta_1^{t+1})} \left((1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \left\| e^{[k,t]} \right\|_2 \right)^2. \end{aligned} \tag{109}$$

Since the weights $(1-\beta_1)\beta_1^{t-k}$ are nonnegative and sum to $1-\beta_1^{t+1}$, weighted Cauchy–Schwarz gives

$$\begin{aligned} \left((1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \left\| e^{[k,t]} \right\|_2 \right)^2 &\leq \left((1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \right) \left((1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \left\| e^{[k,t]} \right\|_2^2 \right) \\ &= (1-\beta_1^{t+1})(1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \left\| e^{[k,t]} \right\|_2^2. \end{aligned} \tag{110}$$

We substitute (110) into (109) and then insert the resulting bound into (108), yielding

$$\begin{aligned}&\langle \nabla \mathcal{J}(\theta^{[t]}), P^{[t]} m^{[t]} \rangle \\ &\geq \hat{\eta} EU \left(\frac{1-\beta_1^{t+1}}{2(\hat{\eta} EU G_\infty + \varepsilon_s)} \left\| \nabla \mathcal{J}(\theta^{[t]}) \right\|_2^2 - \frac{\hat{\eta} EU G_\infty + \varepsilon_s}{2\varepsilon_s^2} (1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \left\| e^{[k,t]} \right\|_2^2 \right).\end{aligned} \tag{111}$$

We now take expectations and control the weighted error term using (107). With $k' = t-k$, we have

$$(1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} (t-k)^2 \leq (1-\beta_1) \sum_{k'=0}^{\infty} \beta_1^{k'} k'^2 = \frac{\beta_1(1+\beta_1)}{(1-\beta_1)^2}, \tag{112}$$

combining which with $(1-\beta_1)\sum_{k=0}^{t} \beta_1^{t-k} = 1 - \beta_1^{t+1} \leq 1$ gives

$$\begin{aligned}&(1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} \mathbb{E}\left[\left\| e^{[k,t]} \right\|_2^2 \right] \\ \leq &\frac{2L^2\eta^2\hat{\eta}^2 E^2 U^2 G_2^2}{\varepsilon_s^2} (1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} (t-k)^2 + 8(1-\beta_1^{t+1}) C_{\mathrm{FGRPO}} \\ \leq &\frac{2L^2\eta^2\hat{\eta}^2 E^2 U^2 G_2^2 \beta_1(1+\beta_1)}{\varepsilon_s^2 (1-\beta_1)^2} + 8C_{\mathrm{FGRPO}}.\end{aligned}$$

Applying this estimate to the expectation of (111) and using $1 - \beta_1^{t+1} \geq 1 - \beta_1$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\langle \nabla \mathcal{J}(\theta^{[t]}), P^{[t]} m^{[t]} \right\rangle \right] \geq &\frac{\hat{\eta} EU(1-\beta_1)}{2(\hat{\eta} EU G_\infty + \varepsilon_s)} \mathbb{E}\left[\left\| \nabla \mathcal{J}(\theta^{[t]}) \right\|_2^2 \right] \\ &- \frac{L^2\eta^2\hat{\eta}^3 E^3 U^3 G_2^2 (\hat{\eta} EU G_\infty + \varepsilon_s) \beta_1(1+\beta_1)}{\varepsilon_s^4 (1-\beta_1)^2} \\ &- \frac{4\hat{\eta} EU(\hat{\eta} EU G_\infty + \varepsilon_s)}{\varepsilon_s^2} C_{\mathrm{FGRPO}}.\end{aligned} \tag{113}$$

To translate this alignment bound into an objective improvement bound, we use L-smoothness and the server update $\theta^{[t+1]} = \theta^{[t]} + \eta P^{[t]} m^{[t]}$:

$$\begin{aligned}\mathcal{J}(\theta^{[t+1]}) &\geq \mathcal{J}(\theta^{[t]}) + \eta \left\langle \nabla \mathcal{J}(\theta^{[t]}), P^{[t]} m^{[t]} \right\rangle - \frac{L\eta^2}{2} \left\| P^{[t]} m^{[t]} \right\|_2^2 \\ &\geq \mathcal{J}(\theta^{[t]}) + \eta \left\langle \nabla \mathcal{J}(\theta^{[t]}), P^{[t]} m^{[t]} \right\rangle - \frac{L\eta^2\hat{\eta}^2 E^2 U^2 G_2^2}{2\varepsilon_s^2},\end{aligned} \tag{114}$$

where the second inequality follows from $\|P^{[t]} m^{[t]}\|_2 \leq \hat{\eta} EU G_2 / \varepsilon_s$ in Lemma A.6.

Taking expectations in (114) and substituting (113) bounds the expected objective improvement from below:

$$\begin{aligned}&\frac{\eta\hat{\eta} EU(1-\beta_1)}{2(\hat{\eta} EU G_\infty + \varepsilon_s)} \mathbb{E}\left[\left\| \nabla \mathcal{J}(\theta^{[t]}) \right\|_2^2 \right] \\ \leq &\mathbb{E}\left[\mathcal{J}(\theta^{[t+1]}) - \mathcal{J}(\theta^{[t]}) \right] \\ &+ \frac{4\eta\hat{\eta} EU(\hat{\eta} EU G_\infty + \varepsilon_s)}{\varepsilon_s^2} C_{\mathrm{FGRPO}} \\ &+ \frac{L^2\eta^3\hat{\eta}^3 E^3 U^3 G_2^2 (\hat{\eta} EU G_\infty + \varepsilon_s) \beta_1(1+\beta_1)}{\varepsilon_s^4 (1-\beta_1)^2} + \frac{L\eta^2\hat{\eta}^2 E^2 U^2 G_2^2}{2\varepsilon_s^2}.\end{aligned} \tag{115}$$

Summing (115) over $t = 0, \ldots, T-1$, the objective differences telescope:

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\mathcal{J}(\theta^{[t+1]}) - \mathcal{J}(\theta^{[t]}) \right] = \mathbb{E}\mathcal{J}(\theta^{[T]}) - \mathcal{J}(\theta^{[0]}) \leq J^* - \mathcal{J}(\theta^{[0]}).$$

The inequality follows from the definition of J^* as an upper bound on the objective. Each of the three error bounds is independent of t, so summing it over the T rounds contributes a factor of T. Hence, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{J}(\theta^{[t]})\|_2^2 \le{}& \frac{2(J^*-\mathcal{J}(\theta^{[0]}))(\hat\eta EUG_\infty+\varepsilon_s)}{\eta\hat\eta EU(1-\beta_1)T}\\ &+\frac{8(\hat\eta EUG_\infty+\varepsilon_s)^2}{\varepsilon_s^2(1-\beta_1)}C_{\mathrm{FGRPO}}\\ &+\frac{2L^2\eta^2\hat\eta^2E^2U^2G_2^2(\hat\eta EUG_\infty+\varepsilon_s)^2\beta_1(1+\beta_1)}{\varepsilon_s^4(1-\beta_1)^3}\\ &+\frac{L\eta\hat\eta EUG_2^2(\hat\eta EUG_\infty+\varepsilon_s)}{\varepsilon_s^2(1-\beta_1)}.\end{aligned}\tag{116}$$

The above inequality is established for a general horizon $T\ge 1$. We now use the step sizes in (99). Write $G_2(T)$ and $G_\infty(T)$ for the bounds in (32) evaluated at $\hat\eta=(\overline{L}EUT^{1/4})^{-1}$, and define

$$C(T)\triangleq(C_{\mathrm{RPG}}+1)\kappa_{\mathrm{var}}^2+(1+C_{\mathrm{RPG}}/E)\,(G_2(T)+\ell_{\max}r_{\max}^*S_2)^2.\tag{117}$$

Only $G_2(T)$ depends on the horizon in (117), through $\hat\eta(T)$ in (31); all other quantities are fixed. For the sharper oracle bound used below, define

$$\mathcal{E}(T)\triangleq(C_{\mathrm{RPG}}-1)\kappa_{\mathrm{var}}^2+\delta_G(T)^2+\frac{C_{\mathrm{RPG}}\nu_G^2(T)}{E}.\tag{118}$$

The oracle estimates in (64) and $C_{\mathrm{RPG}}-1\le C_{\mathrm{RPG}}+1$ imply $\mathcal{E}(T)\le C(T)$. Substitution into (91) gives

$$C_{\mathrm{FGRPO}}\le\frac{L^2G_2(T)^2}{3\overline{L}^2\sqrt{T}}+C(T).$$

Substituting the chosen step sizes and this residual bound into (116) therefore yields

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{J}(\theta^{[t]})\|_2^2\\ &\le\frac{1}{\sqrt{T}}\Bigg(\frac{2\overline{L}(J^*-\mathcal{J}(\theta^{[0]}))\big(\varepsilon_s+G_\infty(T)/(\overline{L}T^{1/4})\big)}{1-\beta_1}\\ &\qquad+\frac{8L^2G_2(T)^2\big(\varepsilon_s+G_\infty(T)/(\overline{L}T^{1/4})\big)^2}{3\varepsilon_s^2\overline{L}^2(1-\beta_1)}\\ &\qquad+\frac{LG_2(T)^2\big(\varepsilon_s+G_\infty(T)/(\overline{L}T^{1/4})\big)}{\varepsilon_s^2\overline{L}(1-\beta_1)}\Bigg)\\ &\quad+\frac{2L^2G_2(T)^2\beta_1(1+\beta_1)\big(\varepsilon_s+G_\infty(T)/(\overline{L}T^{1/4})\big)^2}{\varepsilon_s^4\overline{L}^2(1-\beta_1)^3T}\\ &\quad+\frac{8\big(\varepsilon_s+G_\infty(T)/(\overline{L}T^{1/4})\big)^2}{\varepsilon_s^2(1-\beta_1)}C(T).\end{aligned}\tag{119}$$

This inequality applies to every horizon for which the assumptions hold, without requiring $T\ge T_0$.

The constants in this general bound are uniformly bounded as T varies. Indeed, $\hat\eta(T)\le 1/(\overline{L}EU)$. Evaluate the finite recursion (31) at this fixed upper step size. Since the recursion is nondecreasing in both its preceding radius and its step size, its resulting values bound every $R_u(T)$ independently of T. The fixed reuse count U and primitive bounds, including L_{ref}, then give uniform bounds on $G_2(T)$ and $G_\infty(T)$. Eq. (99) then gives $\hat\eta EUG_2(T)=\mathcal{O}(T^{-1/4})$ and $\hat\eta EUG_\infty(T)=\mathcal{O}(T^{-1/4})$, so $\varepsilon_s+\hat\eta EUG_\infty(T)=\mathcal{O}(1)$. In particular,

$$\begin{aligned}\frac{1}{\eta\hat\eta EUT}&=\frac{\overline{L}}{\sqrt{T}}, & C_{\mathrm{FGRPO}}&\le C(T)+\mathcal{O}(T^{-1/2}),\\ \eta^2\hat\eta^2E^2U^2G_2(T)^2&=\mathcal{O}(T^{-1}), & \eta\hat\eta EUG_2(T)^2&=\mathcal{O}(T^{-1/2}).\end{aligned}\tag{120}$$

Thus the general bound has the form $\mathcal{O}(T^{-1/2}) + \mathcal{O}(C(T))$, before imposing the threshold needed for the simpler explicit constants in the statement.

We next examine how the coarse bound $C(T)$ depends on the horizon T. Its monotonicity does not imply monotonicity of the actual oracle residual $\mathcal{E}(T)$ or of the observed gradient norms. The right-hand side of (31) is nondecreasing in both R_u and $\hat{\eta}$. Since $\hat{\eta}(T)$ decreases with T, induction from $R_0(T) = 0$ shows that $R_u(T)$ is nonincreasing in T for each fixed u. By their definitions, $\rho_{\max}(T)$, $G_2(T)$, and $C(T)$ are therefore also nonincreasing in T. Moreover, when $U > 1$, $S_2 > 0$, and $A_{\max} > 0$, the recursion is strictly increasing in the step size. Thus $R_{U-1}(T)$, and consequently $C(T)$, are strictly decreasing in T.

If $S_2 = 0$, then $R_u(T) = 0$, $\rho_{\max}(T) = 1$, and $G_2(T) = 0$, so the bounds below hold directly. For $S_2 > 0$, set $\zeta_u = \exp(-S_2 R_u(T))$. Since $R_u(T) \geq 0$, we have $0 < \zeta_u \leq 1$. Applying $e^{-x} \geq 1 - x$ gives

$$\begin{aligned}\zeta_{u+1} &= \zeta_u \exp\left(-\hat{\eta} S_2^2 \left(A_{\max}/\zeta_u + \beta \exp(S_2 L_{\mathrm{ref}})\right)\right) \\ &\geq \zeta_u - \hat{\eta} S_2^2 \left(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}})\zeta_u\right) \\ &\geq \zeta_u - \hat{\eta} S_2^2 \left(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}})\right).\end{aligned}$$

Induction from $\zeta_0 = 1$ and the step-size choice in (99) yield

$$\begin{aligned}\zeta_{U-1} &\geq 1 - (U-1)\hat{\eta} S_2^2 \left(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}})\right) \\ &= 1 - \frac{(U-1) S_2^2 (A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}))}{\overline{L} E U T^{1/4}}.\end{aligned}$$

For $S_2 > 0$, this lower bound is positive whenever

$$\overline{L} E U T^{1/4} > (U-1) S_2^2 (A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}})).$$

Under this condition, the definitions of ζ_{U-1} and $\rho_{\max}(T)$ give

$$\begin{aligned}0 \leq R_{U-1}(T) &\leq -\frac{1}{S_2} \log\left(1 - \frac{(U-1) S_2^2 (A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}))}{\overline{L} E U T^{1/4}}\right), \\ 1 \leq \rho_{\max}(T) &\leq \frac{\overline{L} E U T^{1/4}}{\overline{L} E U T^{1/4} - (U-1) S_2^2 (A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}))}.\end{aligned} \tag{121}$$

Substitution into (32) gives

$$\begin{aligned}0 \leq G_2(T) - S_2(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}})) &= S_2 A_{\max}(\rho_{\max}(T) - 1) \\ &\leq \frac{(U-1) S_2^3 A_{\max}(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}))}{\overline{L} E U T^{1/4} - (U-1) S_2^2 (A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}))} \\ &= \mathcal{O}(T^{-1/4}).\end{aligned} \tag{122}$$

Consequently, the coarse residual bound $C(T)$ decreases to

$$C_\infty \triangleq (C_{\mathrm{RPG}} + 1)\kappa_{\mathrm{var}}^2 + (1 + C_{\mathrm{RPG}}/E)\, S_2^2 (A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}) + \ell_{\max} r_{\max}^*)^2. \tag{123}$$

Specifically, with $G_2(T) - S_2(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}})) \geq 0$,

$$\begin{aligned}&C(T) - C_\infty \\ =&(1 + C_{\mathrm{RPG}}/E)\left(G_2(T) - S_2(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}))\right) \\ &\left(2S_2(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}) + \ell_{\max} r_{\max}^*) + (G_2(T) - S_2(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}})))\right) \\ =&\mathcal{O}(T^{-1/4}).\end{aligned} \tag{124}$$

For $U = 1$, $R_{U-1}(T) = R_0 = 0$, so $\rho_{\max}(T) = 1$, $G_2(T) = S_2(A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}))$, and $C(T) = C_\infty$ for every horizon. The same equality holds when $S_2 = 0$. Otherwise, the original $C(T)$ continues to vary with T even though it has a constant limit.

Finally, impose $T \geq T_0$. By (98), this ensures

$$\frac{(U-1) S_2^2 (A_{\max} + \beta \exp(S_2 L_{\mathrm{ref}}))}{\overline{L} E U T^{1/4}} \leq \frac{1}{2},$$

so Eq. (121) gives $\rho_{\max}(T) \le 2$. The same ratio bound holds directly when $S_2 = 0$. Therefore,

$$G_2(T) \le G_2^*, \qquad G_\infty(T) \le G_\infty^*.$$

By (117) and (102), these bounds also imply

$$\frac{8(\varepsilon_s + G_\infty^*/\overline{L})^2}{\varepsilon_s^2(1-\beta_1)} C(T) \le C_{\mathrm{res}}.$$

Moreover, $T \ge 1$ and (99) give

$$\hat{\eta} E U G_\infty(T) = \frac{G_\infty(T)}{\overline{L} T^{1/4}} \le \frac{G_\infty^*}{\overline{L}}, \qquad \frac{1}{\varepsilon_s + \hat{\eta} E U G_\infty(T)} \ge \frac{1}{\varepsilon_s + G_\infty^*/\overline{L}}.$$

To retain the sharper oracle bound, substitute the exact decomposition

$$C_{\mathrm{FGRPO}} = \frac{L^2 G_2(T)^2}{3\overline{L}^2\sqrt{T}} + \mathcal{E}(T)$$

into (116). The algebra leading to (119) is unchanged, with $C(T)$ replaced by $\mathcal{E}(T)$. Substituting $G_2(T) \le G_2^*$ and $G_\infty(T) \le G_\infty^*$ and using $T^{-1} \le T^{-1/2}$ gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla \mathcal{J}(\theta^{[t]})\right\|_2^2\right] \le \frac{C_{\mathrm{tr}}}{\sqrt{T}} + \frac{8(\varepsilon_s + G_\infty^*/\overline{L})^2}{\varepsilon_s^2(1-\beta_1)} \mathcal{E}(T), \tag{125}$$

where C_{tr} collects the transient terms. To explicitly justify the $\mathcal{O}(1)$ residual in (24), note that (64) and $G_2(T) \le G_2^*$ give, for every $T \ge T_0$,

$$\delta_G(T)^2 \le (G_2^* + \ell_{\max} r_{\max}^* S_2)^2, \qquad \nu_G^2(T) \le (G_2^* + \ell_{\max} r_{\max}^* S_2)^2.$$

Substituting these bounds into (118) yields

$$\begin{aligned} 0 \le \mathcal{E}(T) &= (C_{\mathrm{RPG}} - 1)\kappa_{\mathrm{var}}^2 + \delta_G(T)^2 + \frac{C_{\mathrm{RPG}} \nu_G^2(T)}{E} \\ &\le (C_{\mathrm{RPG}} - 1)\kappa_{\mathrm{var}}^2 + \left(1 + \frac{C_{\mathrm{RPG}}}{E}\right)(G_2^* + \ell_{\max} r_{\max}^* S_2)^2. \end{aligned}$$

All quantities on the last line are finite and independent of T: the theorem fixes E, U, K, B, the initialization, and the problem and optimizer constants, and G_2^*, G_∞^* are defined independently of T in (100). Moreover, $\varepsilon_s > 0$ and $\beta_1 < 1$ make the residual multiplier in (125) finite. Using $C_{\mathrm{RPG}} - 1 \le C_{\mathrm{RPG}} + 1$ and the definition of C_{res}, we therefore obtain the uniform bound

$$0 \le \frac{8(\varepsilon_s + G_\infty^*/\overline{L})^2}{\varepsilon_s^2(1-\beta_1)} \mathcal{E}(T) \le C_{\mathrm{res}} < \infty, \qquad T \ge T_0.$$

Since C_{res} does not depend on T, this is precisely an $\mathcal{O}(1)$ bound on the residual contribution as $T \to \infty$. Likewise, C_{tr} is finite and independent of T, so $C_{\mathrm{tr}}/\sqrt{T} = \mathcal{O}(T^{-1/2})$. Together with (125), these estimates prove both (101) and (24). The $\mathcal{O}(1)$ notation asserts uniform boundedness only; it does not assert that the residual vanishes or has a positive limit. □

Theorem A.1 bounds the average expected squared gradient norm by a term decreasing as $T^{-1/2}$ plus a residual. The decreasing term includes local drift; the residual bound covers RPG weighting, GRPO bias, and sampling noise, which additional rounds alone need not remove. The guarantee applies to the average over communication rounds, or equivalently to a uniformly selected round, rather than specifically to the final model.

RPG aggregation is designed to mitigate the effects of heterogeneous client data by prioritizing clients with larger relative gains; it does not make their data distributions or population gradients identical. Because the target objective $\mathcal{J} = \sum_i p_i \mathcal{J}_i$ uses fixed weights, RPG's adaptive weights $\omega_i^{[t]}$ may also change the aggregate direction relative to $\nabla \mathcal{J}$. The term $(C_{\mathrm{RPG}} - 1)\kappa_{\mathrm{var}}^2$ in (118) bounds this possible discrepancy. The actual RPG residual vanishes if $\omega_i^{[t]} = p_i$ for every client or if all client population gradients agree. Positive temperature and volatility floors limit weight concentration

Table 2: Per-round communication overhead of FGRPO for each client across different backbone models.

Model	Qwen2.5-3B	Qwen3-4B	Qwen2.5-7B	Llama-3.2-11B
Communication overhead (MB)	228.45	252.07	308.06	400.08

through C_{RPG}. The theorem thus quantifies a possible cost of adaptive weighting; whether RPG improves learning under heterogeneity is assessed empirically.

The remaining terms in $\mathcal{E}(T)$ measure GRPO bias and sampling noise. The bias $\delta_G(T)^2$ reflects a possible difference between the mean GRPO update and the gradient of the reward objective, owing to group-relative advantages, clipping, response-length normalization, and KL regularization. The noise term $C_{\mathrm{RPG}}\nu_G^2(T)/E$ reflects averaging over fresh rollout cycles and, when $U = 1$, is at most $C_{\mathrm{RPG}}G_2(T)^2/(BE)$ by (65). Reusing a sampled block does not add independent data, and noise cannot simply be discarded because RPG weights and the Adam preconditioner depend on the samples. The refined bound (125) combines a $T^{-1/2}$ transient with a term proportional to $\mathcal{E}(T)$. The standing assumptions bound $\mathcal{E}(T)$ but do not ensure it decreases; the expanded C_{res} is a conservative uniform upper bound. As discussed in Sec. 3.3, residual terms also appear in convergence bounds for state-of-the-art FedRL methods with critics (Yang et al., 2024; Xie & Song, 2025). Establishing decay with T may require stronger assumptions or sharper bounds on the underlying discrepancies. A systematic characterization of the additional conditions needed is beyond the scope of this paper.

B More Information about Experiment Settings

B.1 Implementation

To improve computational and communication efficiency, we incorporate Low-Rank Adaptation (LoRA) (Hu et al., 2022) into FGRPO, freezing the pre-trained backbone weights and training only low-rank adapters. For a linear layer with frozen weight $\mathbf{W}_0$, LoRA uses the effective weight $\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W}$ during the forward pass, where $\Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{B}\mathbf{A}$, $\mathbf{A}$ and $\mathbf{B}$ are trainable low-rank matrices, r is the LoRA rank, and α is the scaling factor. Accordingly, each client $i \in \mathcal{N}$ in Algorithm 1 optimizes only its local adapter parameters $\theta_i = \{\mathbf{A}_i, \mathbf{B}_i\}$ throughout all communication rounds, with the local direction $g_i^{[t,e,u]}$ computed only with respect to these parameters. The server communicates and aggregates the adapter matrices, substantially reducing the number of trainable and transmitted parameters, GPU memory usage, and computational and communication costs.

We implement LoRA under our FGRPO framework by restricting both local optimization and server-client communication to the adapter parameters. Specifically, LoRA modules are inserted into the target linear projection layers of each decoder layer in the language-model backbone, including the self-attention projections (`q_proj`, `k_proj`, `v_proj`, and `o_proj`) and the MLP projections (`gate_proj`, `up_proj`, and `down_proj`). We set the LoRA rank to $r = 32$ and the scaling factor to $\alpha = 128$. At the beginning of each communication round t, the server broadcasts only the global adapter parameters $\theta^{[t]} = \{\mathbf{A}^{[t]}, \mathbf{B}^{[t]}\}$ to the clients, while the backbone weights $\mathbf{W}_0$ remain fixed throughout training. Each client then performs local GRPO updates on its private data, where gradients are computed only with respect to the LoRA adapters. After local training, client i uploads its updated adapters $\theta_i^{[t+1]} = \{\mathbf{A}_i^{[t+1]}, \mathbf{B}_i^{[t+1]}\}$ together with its RPG value $h_i^{[t]}$, computed locally from the round-level mean reward $\bar{r}_i^{[t]}$. The server applies our RPG-based adaptive weighting mechanism to these uploaded adapters and aggregates them to form the next global adapter.

We report the per-round communication overhead of the LoRA-based implementation in Table 2. Since LoRA freezes the backbone and updates only low-rank adapter parameters, it substantially reduces the number of transmitted parameters. The communication overhead remains moderate across model scales, ranging from 228.45 MB to 400.08 MB. Specifically, the overhead is 228.45 MB, 252.07 MB, 308.06 MB, and 400.08 MB for the 3B, 4B, 7B, and 11B models, respectively. These results show that even when scaling to larger backbones, FGRPO communicates only a compact set of trainable adapter parameters, making federated reasoning-policy optimization practical under limited communication resources.

B.2 Datasets and Baselines

We conduct our extensive experiments on the following two datasets:

- **Multimodal-Open-R1-8k-Verified** (EvolvingLMMs-Lab, 2025) consists of 8,000+ high-quality samples focusing on diverse mathematical disciplines, including algebra, probability, and functional analysis. It requires models to perform multi-step deduction grounded in various visual contexts such as statistical charts and functional plots. In our framework, we adopt an outcome-based verification protocol to evaluate the model's ability to synthesize coherent, long-chain reasoning paths across broad logical domains, focusing on the accuracy of the final answer.

- **GEOQA-8k** (Chen et al., 2025a) is designed for specialized geometric reasoning and comprises 8,000+ problems pairing 2D geometric diagrams with natural language descriptions. It emphasizes spatial reasoning and geometric constraints, requiring models to derive precise numerical solutions through joint spatial-logical synthesis. Following the same outcome-based verification protocol, the framework is assessed on its capacity to arrive at correct answers based on axiomatic theorems without intermediate process-based supervision.

For the OpenR1 dataset, we stratify samples into three difficulty tiers (*simple*, *medium*, and *hard*) based on equal tertiles of reasoning trace length. We use trace length as a metric to evaluate task complexity, as longer traces typically involve more logical transitions and self-correction steps. To simulate data heterogeneity, we allocate these tiers to clients using a Dirichlet distribution parameterized by μ, so that different clients are exposed to varying levels of reasoning difficulty. We examine different levels of data heterogeneity by varying $\mu \in \{0.05, 1.0\}$, corresponding to *highly* and *moderately* non-IID regimes, respectively, along with a *uniform* baseline with $\mu \to \infty$.

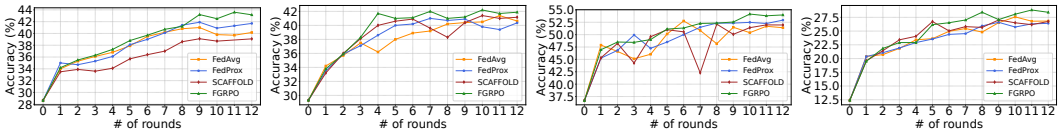
For the GEOQA dataset, we model domain heterogeneity by partitioning samples according to topological complexity and geometric primitives (e.g., *points*, *lines*, *circles*, and *polygons*). Using the same Dirichlet-based allocation, we assign clients to specialize in disjoint visual concepts, such as circular reasoning versus polygonal construction. Unlike the complexity-driven imbalance in OpenR1, this setup evaluates the model's ability to aggregate diverse feature representations and generalize across structurally distinct geometric subdomains without incurring catastrophic forgetting.

In this paper, we compare FGRPO with the following three representative FL baselines:

- **FedAvg** (McMahan et al., 2017) is the classical FL algorithm, which performs iterative model averaging across decentralized clients. In each communication round, each client performs local GRPO optimization on its private dataset, and the server aggregates the resulting local models via a weighted average. Specifically, the aggregation weights are proportional to the local data volume, i.e., $\omega_i = \frac{D_i}{\sum_{i'=1}^{N} D_{i'}}$, where D_i denotes the size of client i's dataset. This procedure enables collaborative model training without sharing raw data, while naturally emphasizing clients with larger datasets.

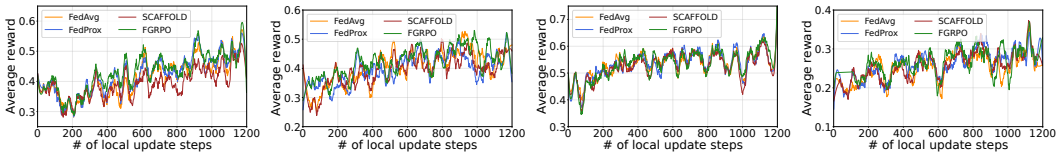
- **FedProx** (Li et al., 2020) is a heterogeneity-aware extension of FedAvg designed to improve federated optimization under non-IID data. In the FGRPO setting, we adapt FedProx by subtracting the proximal penalty $\frac{\mu_{\text{prox}}}{2}\|\theta_i - \theta^{[t]}\|_2^2$, with $\mu_{\text{prox}} \geq 0$, from each client's maximized local GRPO surrogate. Its contribution to the ascent direction is $-\mu_{\text{prox}}(\theta_i - \theta^{[t]})$, which pulls the local parameters toward the broadcast model.

- **SCAFFOLD** (Karimireddy et al., 2020) is a variance-reduction-based federated optimization method that explicitly addresses client drift under non-IID data distributions. Specifically, SCAFFOLD equips each client i with control variates that are incorporated into its local direction $g_i^{[t,e,u]}$, thereby correcting the discrepancy between local and global update directions. These control variates estimate the deviation between the client-specific gradient and the global objective, enabling each local update to better align with the global optimization trajectory and significantly reducing the variance induced by heterogeneous data.



(a) 4B model on OpenR1 (b) 11B model on OpenR1 (c) 4B model on GEOQA (d) 11B model on GEOQA

Figure 3: Test accuracy convergence trajectories of the different models (Qwen3-4B and Llama-3.2-11B) on OpenR1 and GEOQA datasets.



(a) 4B model on OpenR1 (b) 11B model on OpenR1 (c) 4B model on GEOQA (d) 11B model on GEOQA

Figure 4: Average reward trajectories of the different models (Qwen3-4B and Llama-3.2-11B) on OpenR1 and GEOQA datasets.

C Supplementary Experiment Results

We first present complementary comparison results in Sec.C.1. We then analyze the impact of the number of clients in Sec.C.2, followed by the effect of data heterogeneity in Sec.C.3. Next, we examine the extension of our framework to GRPO-family variants with RPG ablations in Sec.C.4, and conduct hyperparameter sensitivity analysis in Sec.C.5. Finally, we report the resource consumption of FGRPO in Sec.C.6.

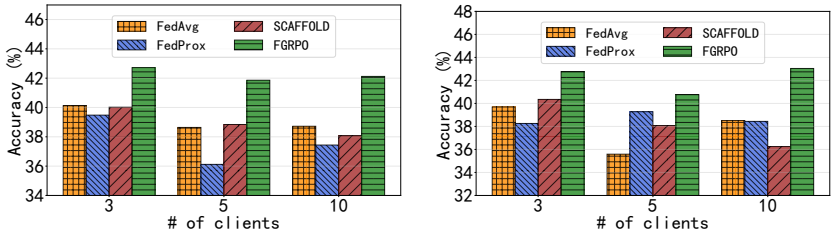
C.1 Additional Comparison Results

In this section, we provide additional comparison results on Qwen3-4B and Llama-3.2-11B models to further complement the results shown in Figs. 1–2. As demonstrated in Fig. 3, for Qwen3-4B model, FGRPO exhibits a clear upward trajectory and reaches the highest final accuracy on both OpenR1 and GEOQA. In contrast, the baseline methods either plateau earlier or show more noticeable fluctuations in later communication rounds. For Llama-3.2-11B model, FGRPO also maintains a stable advantage in the later stages of training, especially on GEOQA, where it consistently stays above the other federated baselines after the middle communication rounds. These observations indicate that RPG-based aggregation can provide stable optimization benefits across different models.

The reward trajectories in Fig. 4 are consistent with the accuracy curves. On both Qwen3-4B and Llama-3.2-11B models, FGRPO generally achieves more favorable average reward trajectories across local update steps. This shows that FGRPO not only improves final evaluation accuracy, but also leads to more effective reinforcement learning dynamics during federated training. By aggregating clients according to relative performance gains rather than absolute reward magnitudes, FGRPO better captures meaningful local progress under heterogeneous reward distributions.

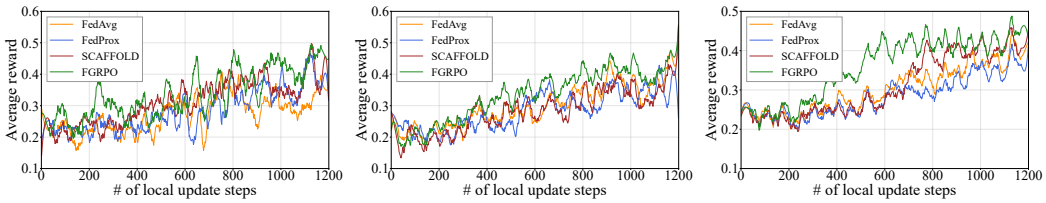
C.2 Impact of the Number of Clients

We evaluate the test accuracy of different methods under varying numbers of clients ($N \in \{3, 5, 10\}$) using the Qwen2.5-3B model, as shown in Fig. 5. The results show that FGRPO consistently outperforms all baselines, maintaining an accuracy above 42.12% on OpenR1 even at $N = 10$, substantially higher than the 37.44% achieved by FedProx, demonstrating strong scalability to larger client populations. Moreover, the performance gap between FGRPO and the baselines widens as the number of clients increases. On GEOQA, the accuracy gap between FGRPO and SCAFFOLD grows from 2.42 percentage points at $N = 3$ to 6.79 percentage points at $N = 10$, suggesting that FGRPO remains effective when aggregating updates from more clients. A similar trend is observed against FedProx, where the margin improves from 4.51 to 4.60 percentage points. Overall, these results



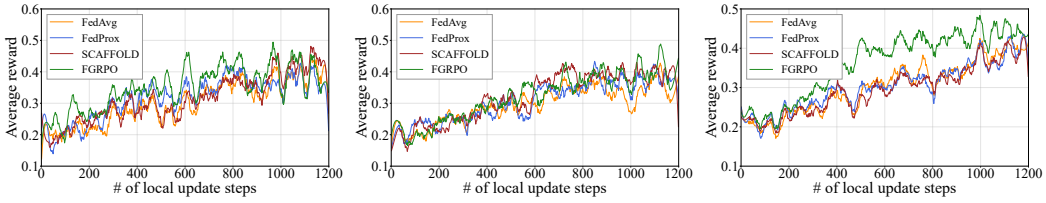
(a) OpenR1 (b) GEOQA

Figure 5: Test accuracy under different numbers of clients.



(a) 3 clients (b) 5 clients (c) 10 clients

Figure 6: Average reward trajectories under varying numbers of clients on OpenR1 dataset.

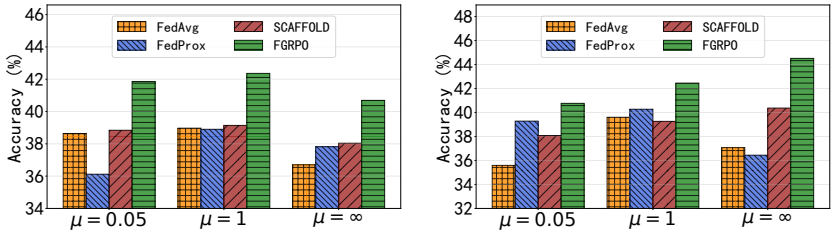


(a) 3 clients (b) 5 clients (c) 10 clients

Figure 7: Average reward trajectories under varying numbers of clients on GEOQA datasets.

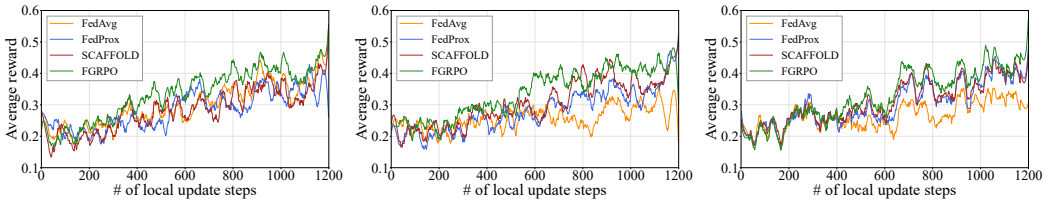
demonstrate that RPG-based aggregation enhances learning performance across varying numbers of clients.

To further analyze the effect of client population size, we report the average reward trajectories in Fig. 6 and Fig. 7. On both datasets, all methods show a generally increasing reward trend as local training proceeds, while the reward curves become relatively smoother when the number of clients increases from $N = 3$ to $N = 10$. This suggests that aggregating feedback from a larger client population can partially reduce the stochastic fluctuations caused by individual local updates. Compared with FedAvg, FedProx, and SCAFFOLD, FGRPO consistently maintains higher average reward trajectories across different client population sizes. When the federation is small ($N = 3$), the training dynamics are more volatile because each client has a larger influence on the global update. Even in this setting, FGRPO achieves higher reward levels and shows stronger robustness to local reward noise. As the number of clients increases to $N = 5$ and $N = 10$, FGRPO further exhibits a more stable upward trend and sustains a clear reward advantage over the baselines on both Open-R1 and GEOQA. These results indicate that the benefit of FGRPO is not limited to a specific federation size. While increasing the number of clients can improve the stability of federated aggregation, the performance gap between FGRPO and the baselines shows that client population size alone is insufficient. The proposed RPG-based aggregation remains important for identifying clients with meaningful relative reward improvements and for mitigating the negative effect of heterogeneous reward distribution during the federated training.



(a) OpenR1 (b) GEOQA

Figure 8: Test accuracy under varying data heterogeneity on OpenR1 and GEOQA datasets.



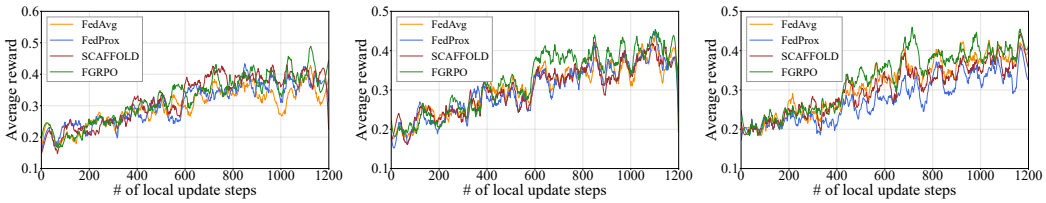
(a) $\mu = 0.05$ (b) $\mu = 1$ (c) $\mu = \infty$

Figure 9: Average reward trajectories of different algorithms under varying non-IID data settings on OpenR1 dataset.

C.3 Impact of Data Heterogeneity Levels

We further evaluate the performance of different algorithms under varying levels of data heterogeneity using the Qwen2.5-3B model. Specifically, we simulate non-IID data distributions across five clients by adjusting the Dirichlet parameter μ, where smaller μ leads to more skewed client partitions, and $\mu \to \infty$ corresponds to a uniform distribution. Fig. 8 presents the performance of different methods under three representative settings: *high* heterogeneity ($\mu = 0.05$), *moderate* heterogeneity ($\mu = 1.0$), and *low* heterogeneity ($\mu \to \infty$). The results show that FGRPO consistently achieves the best overall performance on both OpenR1 and GEOQA under all heterogeneity levels. Under the most challenging highly non-IID setting ($\mu = 0.05$), FGRPO reaches 41.86% accuracy on OpenR1, outperforming FedAvg, FedProx, and SCAFFOLD by 3.22, 5.74, and 3.02 percentage points, respectively. A similar advantage can be observed on GEOQA, where FGRPO achieves 40.76% accuracy and surpasses the strongest baseline by 1.48 percentage points. These results demonstrate that FGRPO remains robust even when local clients are exposed to substantially different reasoning difficulties or visual concept distributions. As the data distribution becomes less heterogeneous, FGRPO continues to maintain a clear advantage over the baselines. In the medium-heterogeneity setting ($\mu = 1.0$), FGRPO improves the best competing baseline by 3.22 percentage points on OpenR1 and 2.18 percentage points on GEOQA. In the uniform setting ($\mu = \infty$), FGRPO still achieves the highest accuracy, with gains of 2.65 and 4.14 percentage points over the best baseline on OpenR1 and GEOQA, respectively. This indicates that the benefit of FGRPO is not limited to extreme non-IID scenarios; instead, its adaptive aggregation mechanism also improves generalization when client distributions are relatively balanced. Overall, these results confirm that conventional federated optimization methods such as FedAvg, FedProx, and SCAFFOLD are insufficient to address the heterogeneous reward dynamics in the federated environment. By weighting client updates according to RPG, FGRPO can better identify effective local learning trajectories and suppress less reliable updates, leading to more stable and robust performance across different degrees of data heterogeneity.

Fig. 9 and Fig. 10 show the average reward trajectories under different data heterogeneity settings on OpenR1 and GEOQA datasets, respectively. FGRPO consistently demonstrates superior performance, maintaining higher average reward levels. This advantage is most pronounced in the highly non-IID scenario (Fig. 9 (a), 10 (a)), where the highly skewed data distributions cause severe reward fluctuations for the baselines. In this challenging regime, FGRPO's trajectory remains higher than the baselines, confirming that the RPG-based aggregation effectively mitigates the optimization difficulties arising from gradient conflicts and reward distirbution discrepancies.



(a) $\mu = 0.05$ (b) $\mu = 1$ (c) $\mu = \infty$

Figure 10: Average reward trajectories of different algorithms under varying non-IID data settings on GEOQA dataset.

Table 3: Test accuracy (%) of applying the proposed federated framework to different RLVR algorithms on the Open-R1 and GEOQA benchmarks. Results are reported as mean±std.

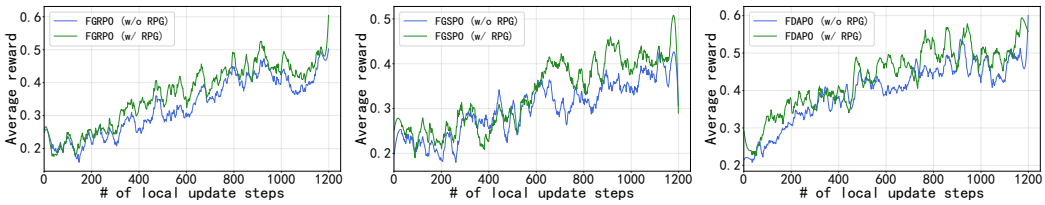
Setup		Open-R1				GEOQA				
Algorithm	**Variant**	**Simple**	**Medium**	**Hard**	**Total**	**Points**	**Lines**	**Circles**	**Polygons**	**Total**
FGRPO	w/ RPG	48.11±0.96	44.14±3.87	33.35±2.19	41.86±1.21	47.62±4.12	45.34±2.64	46.65±1.38	35.88±1.40	40.76±0.74
	w/o RPG	49.01±2.53	42.04±3.44	30.96±2.12	40.66±1.21	43.33±5.16	39.25±1.98	47.22±2.49	33.13±2.06	38.66±1.88
FGSPO	w/ RPG	47.21±0.83	42.76±2.03	32.34±2.19	40.76±0.92	35.24±5.43	38.51±2.88	39.38±0.61	34.50±1.19	36.55±0.84
	w/o RPG	47.21±1.25	41.50±3.39	31.32±3.44	40.00±1.26	34.29±6.86	37.02±2.87	37.09±2.40	31.12±1.22	33.80±0.54
FDAPO	w/ RPG	44.68±1.33	43.12±1.63	38.74±1.58	42.18±0.24	46.67±8.68	52.67±4.70	50.35±1.09	39.13±0.92	44.50±0.96
	w/o RPG	43.84±1.20	42.10±1.89	36.59±2.25	40.84±0.99	36.19±6.82	42.73±6.88	47.14±1.73	34.32±1.50	39.45±1.82

C.4 Extensions to GRPO-based Variants with RPG Ablations

FGRPO is a general federated framework for GRPO-family methods under the paradigm of *reinforcement learning with verifiable rewards* (RLVR), rather than being tied to a specific algorithmic variant. On the client side, it requires a local GRPO-style update and an RPG value computed from round-level rewards, while the server performs RPG-based adaptive aggregation. This design enables recent GRPO extensions, such as DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025), to be seamlessly incorporated, yielding their federated counterparts, **FDAPO** and **FGSPO**. To evaluate the effect of the RPG-based adaptive weighting mechanism, we compare these federated variants (FGRPO, FDAPO, and FGSPO) with their counterparts that use standard data-volume-based aggregation instead of RPG. Specifically, we fine-tune a Qwen2.5-3B model on the Open-R1 and GEOQA datasets using each method, and report the resulting test accuracy in Table 3.

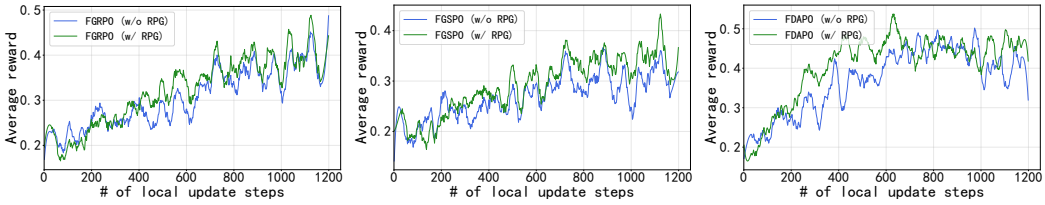
The results show that federated RLVR algorithms equipped with the RPG-based adaptive weighting mechanism consistently outperform their ablated counterparts in terms of total test accuracy. For FGRPO on Open-R1, RPG-based weighting improves test accuracy from 40.66% to 41.86%, a gain of 1.20 percentage points. Similarly, FDAPO achieves a 1.34-percentage-point improvement, increasing accuracy from 40.84% to 42.18%. FGSPO also benefits from RPG, improving from 40.00% to 40.76%. The improvement is more significant on GEOQA. FGRPO improves over its non-RPG variant by 2.10 percentage points, while FDAPO achieves a larger improvement of 5.05 percentage points, increasing the accuracy from 39.45% to 44.50%. FGSPO also obtains a substantial gain of 2.75 percentage points. These results indicate that the RPG-based weighting mechanism provides consistent performance gains across different RLVR algorithms, demonstrating its general effectiveness.

Fig.11 and Fig.12 illustrate the reward trajectories of federated RLVR algorithms with and without our RPG-based weighting mechanism. Across GRPO, GSPO, and DAPO, the RPG-based variants consistently exhibit more stable and favorable reward progression, aligning with the final accuracy improvements reported in Table 3. These results imply that leveraging the concept of RPG can more effectively capture meaningful client-side progress and mitigate the impact of heterogeneous data distribution. These results demonstrate the generality of RPG across different RLVR algorithms. Although the underlying objectives lead to different absolute performance levels, incorporating RPG consistently improves accuracy on both Open-R1 and GEOQA, highlighting its role as a general reward-aware aggregation mechanism for federated RLVR.



(a) FGRPO (b) FGSPO (c) FDAPO

Figure 11: Average reward trajectories of different federated RLVR algorithms on OpenR1 dataset.



(a) FGRPO (b) FGSPO (c) FDAPO

Figure 12: Average reward trajectories of different federated RLVR algorithms on GEOQA dataset.

C.5 Hyperparameter Sensitivity Analysis

We conduct a hyperparameter sensitivity analysis on the Qwen2.5-3B model using the Open-R1 dataset. Our study focuses on six key hyperparameters in FGRPO: the EMA coefficient λ_{base}, the temperature bounds $(\tau_{\min}, \tau_{\max})$, the volatility bounds $(\sigma_{\min}, \sigma_{\max})$, and the annealing coefficient λ_{anneal}. Unless otherwise specified, we adopt the default setting $\lambda_{\text{base}} = 0.8$, $\tau_{\min} = 1.5$, $\tau_{\max} = 2.5$, $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.2$, and $\lambda_{\text{anneal}} = 0.1$.

The coefficient λ_{base} controls the exponential moving average used to smooth the estimation of client-side progress. A larger λ_{base} makes the baseline more sensitive to recent reward changes, while a smaller λ_{base} places more emphasis on historical estimates. As shown in Table 4, the default value $\lambda_{\text{base}} = 0.8$ achieves the best total accuracy of 41.86%. When λ_{base} is reduced to 0.6 or increased to 0.9, the total accuracy becomes 40.98% and 40.82%, respectively. Although $\lambda_{\text{base}} = 0.6$ slightly improves the Hard split, it reduces the Simple and Total performance. This indicates that a moderate EMA coefficient provides a better balance between responsiveness and stability.

The temperature bounds $(\tau_{\min}, \tau_{\max})$ control the range of adaptive aggregation sharpness. A lower temperature makes the aggregation more concentrated on clients with larger relative gains, while a higher temperature produces smoother and more uniform weighting. Table 5 shows that FGRPO is relatively robust to different temperature ranges. The default setting $(\tau_{\min}, \tau_{\max}) = (1.5, 2.5)$ obtains the best total accuracy of 41.86%. Changing the upper bound to 2.0 or 3.0 only slightly decreases the total accuracy to 41.48% and 41.58%, respectively. Interestingly, increasing $\tau_{\max}$ to 3.0 improves the Hard split to 35.33%, but reduces Simple and Medium accuracy, suggesting a trade-off between difficult-example optimization and overall balanced performance. Similarly, varying $\tau_{\min}$ also leads to only moderate performance changes, with total accuracy remaining above 41.12%.

The volatility bounds $(\sigma_{\min}, \sigma_{\max})$ define the clipping range for reward-progress volatility estimation. These bounds prevent the aggregation weights from becoming overly sensitive to unstable reward fluctuations. As shown in Table 6, the default setting $(\sigma_{\min}, \sigma_{\max}) = (0.05, 0.2)$ achieves the best total accuracy of 41.86%. Reducing $\sigma_{\min}$ to 0.03 or increasing it to 0.10 decreases the total accuracy to 40.86% and 40.54%, respectively. Similarly, changing $\sigma_{\max}$ to 0.15 or 0.25 results in total accuracies of 41.04% and 40.90%. These results suggest that properly bounding the volatility estimate is important for stabilizing RPG-based aggregation, especially under non-IID data.

Finally, λ_{anneal} controls the annealing strength of the adaptive aggregation process. It determines how quickly the aggregation behavior changes during training. As shown in Table 7, the default value $\lambda_{\text{anneal}} = 0.1$ achieves the best total accuracy of 41.86%. Reducing it to 0.05 leads to a total

Table 4: Sensitivity analysis of λ_{base} with other hyperparameters fixed to $\tau_{\min} = 1.5$, $\tau_{\max} = 2.5$, $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.2$, and $\lambda_{\text{anneal}} = 0.1$.

Hyperparameter	Simple	Medium	Hard	Total
$\lambda_{\text{base}} = 0.8$	48.11	44.14	33.35	41.86
$\lambda_{\text{base}} = 0.6$	45.05	43.84	34.07	40.98
$\lambda_{\text{base}} = 0.9$	46.79	42.88	32.81	40.82

Table 5: Sensitivity analysis of $\tau_{\min}$ and $\tau_{\max}$ with other hyperparameters fixed to $\lambda_{\text{base}} = 0.8$, $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.2$, and $\lambda_{\text{anneal}} = 0.1$.

Hyperparameter	Simple	Medium	Hard	Total
$\tau_{\min} = 1.5,\ \tau_{\max} = 2.5$	48.11	44.14	33.35	41.86
$\tau_{\min} = 1.5,\ \tau_{\max} = 2.0$	49.85	43.55	31.06	41.48
$\tau_{\min} = 1.5,\ \tau_{\max} = 3.0$	46.85	42.58	35.33	41.58
$\tau_{\min} = 1.0,\ \tau_{\max} = 2.5$	48.23	41.08	34.43	41.24
$\tau_{\min} = 2.0,\ \tau_{\max} = 2.5$	48.41	43.12	31.86	41.12

Table 6: Sensitivity analysis of $\sigma_{\min}$ and $\sigma_{\max}$ with other hyperparameters fixed to $\lambda_{\text{base}} = 0.8$, $\tau_{\min} = 1.5$, $\tau_{\max} = 2.5$, and $\lambda_{\text{anneal}} = 0.1$.

Hyperparameter	Simple	Medium	Hard	Total
$\sigma_{\min} = 0.05,\ \sigma_{\max} = 0.2$	48.11	44.14	33.35	41.86
$\sigma_{\min} = 0.03,\ \sigma_{\max} = 0.2$	49.55	42.70	30.36	40.86
$\sigma_{\min} = 0.10,\ \sigma_{\max} = 0.2$	48.35	41.86	31.44	40.54
$\sigma_{\min} = 0.05,\ \sigma_{\max} = 0.15$	46.97	43.84	32.34	41.04
$\sigma_{\min} = 0.05,\ \sigma_{\max} = 0.25$	48.11	42.94	31.68	40.90

accuracy of 41.24%, while increasing it to 0.15 decreases the total accuracy to 40.76%. Although $\lambda_{\text{anneal}} = 0.15$ slightly improves the Medium split, it significantly hurts the Hard split, indicating that overly aggressive annealing may destabilize optimization for difficult reasoning examples.

Overall, the sensitivity results show that FGRPO is reasonably robust within a practical range of hyperparameter choices. The total accuracy remains around 40.5%–41.9% across different settings, and the default configuration consistently achieves the best overall performance. These results also suggest that the proposed RPG-based aggregation does not rely on a narrowly tuned hyperparameter configuration, while moderate smoothing, bounded volatility estimation, and stable annealing are beneficial for federated reasoning optimization.

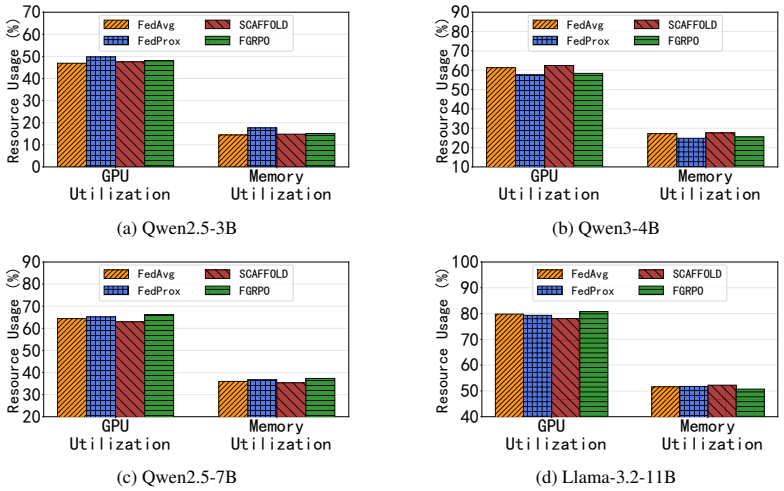
C.6 Resource Consumption

As discussed in Appendix B, the communication overhead per round is mainly determined by the size of the LoRA low-rank matrices. Since all methods use the same LoRA configuration, they incur comparable communication costs. Therefore, we focus on evaluating computational cost by analyzing GPU utilization and GPU memory utilization. Fig. 13 reports these metrics, averaged across the five clients per communication round. With Qwen2.5-3B model (Fig. 13 (a)), all algorithms exhibit comparable resource consumption. FedAvg achieves an averaged GPU utilization of 46.91% and memory utilization of 14.45%, while FedProx and SCAFFOLD consume 49.74%/17.65% and 47.52%/14.83%, respectively. FGRPO records 47.94% GPU utilization and 15.10% memory utilization, remaining very close to FedAvg and SCAFFOLD and lower than FedProx in both metrics. This indicates that the proposed RPG-based aggregation mechanism introduces negligible additional overhead for the 3B model.

A similar trend is observed under the Qwen3-4B model (Fig. 13 (b)), where overall resource utilization increases due to the larger model size. FedAvg and SCAFFOLD show relatively higher averaged GPU utilization of 61.30% and 62.43%, with memory utilization of 27.22% and 27.57%, respectively.

Table 7: Sensitivity analysis of λ_{anneal} with other hyperparameters fixed to $\lambda_{\text{base}} = 0.8$, $\tau_{\min} = 1.5$, $\tau_{\max} = 2.5$, $\sigma_{\min} = 0.05$, and $\sigma_{\max} = 0.2$.

Hyperparameter	Simple	Medium	Hard	Total
$\lambda_{\text{anneal}} = 0.1$	48.11	44.14	33.35	41.86
$\lambda_{\text{anneal}} = 0.05$	49.73	40.96	33.05	41.24
$\lambda_{\text{anneal}} = 0.15$	48.77	44.68	28.86	40.76



(a) Qwen2.5-3B (b) Qwen3-4B

(c) Qwen2.5-7B (d) Llama-3.2-11B

Figure 13: Average GPU and memory utilization across different models.

FedProx consumes 57.62% GPU and 24.82% memory. In comparison, FGRPO achieves 58.37% GPU utilization and 25.53% memory utilization, remaining within the same cost envelope as the baselines and even reducing resource usage compared with FedAvg and SCAFFOLD.

Under the Qwen2.5-7B model (Fig. 13 (c)), the resource demand further increases for all methods. FedAvg, FedProx, and SCAFFOLD achieve averaged GPU utilization of 64.38%, 65.14%, and 62.94%, respectively, with corresponding memory utilization of 35.99%, 36.69%, and 35.32%. The full FGRPO consumes 66.21% GPU and 37.33% memory. Although FGRPO is slightly higher than the baselines in this setting, the difference remains marginal, suggesting that the adaptive aggregation strategy does not introduce significant additional per-round computational or memory cost.

Finally, under the Llama-3.2-11B model (Fig. 13 (d)), all methods reach higher GPU and memory utilization due to the substantially larger model scale. FedAvg, FedProx, and SCAFFOLD consume 79.81%/51.60%, 79.34%/51.76%, and 78.05%/52.25% GPU/memory utilization, respectively. In contrast, FGRPO achieves an averaged GPU utilization of 80.78%, which is slightly higher than the baselines, while maintaining the lowest memory utilization at 50.72%. This demonstrates that FGRPO can scale to larger models without introducing additional memory pressure.

Overall, these results demonstrate that the performance gains of FGRPO are achieved without incurring significant additional computational or memory overhead. Across all model scales, FGRPO remains within the same resource-consumption range as conventional federated baselines, while providing stronger robustness and accuracy under heterogeneous data distributions.

D Comprehensive Literature Survey

D.1 Federated Reinforcement Learning

Federated reinforcement learning (FedRL) aims to leverage the principles of collaborative learning (McMahan et al., 2017) across diverse clients to enhance sample efficiency without compromising raw data or trajectory privacy. Recent theoretical research has focused on establishing rigorous

convergence guarantees under the unique constraints of sequential decision-making. Zhang et al. (2024) provide a fundamental finite-time analysis of on-policy FedRL under data heterogeneity, while Khodadadian et al. (2022) demonstrate that linear speedup is achievable even under the complexities of Markovian sampling. In the offline setting, it is proved in Woo et al. (2024) that a "collaborative single-policy coverage" condition, where the union of client data covers the optimal policy, is sufficient for global optimality. Furthermore, Xiong et al. (2025) highlight that shared representation learning can further accelerate convergence by extracting collaborative features across diverse tasks, while Yang et al. (2024) propose federated natural policy gradient methods to exploit common tasks.

Beyond theoretical convergence, recent works address the practical challenges of system reliability and environmental diversity. Robustness against failures and adversaries has been explored by Fan et al. (2021) and Fang et al. (2025), who provide formal certifications of policy performance under Byzantine disruptions and state perturbations. To handle system-level constraints, Lan et al. (2025) introduce an asynchronous framework that ensures robustness to varying computational speeds, while Jordan et al. (2024) propose a fully decentralized policy gradient algorithm for peer-to-peer topologies. Managing heterogeneous dynamics and architectures also remains a central theme; while Wang et al. (2024) use momentum-based aggregation to stabilize updates across diverse environments, Jiang et al. (2025) introduce FedHPD, utilizing policy distillation to enable collaborative learning between clients with different model structures. Finally, Woo et al. (2023) prove that diverse local distributions can actually reduce global coverage requirements, while Rengarajan et al. (2024) use ensemble-directed models to quantify uncertainty in offline settings.

D.2 GRPO

Group relative policy optimization (GRPO) (Shao et al., 2024) has established a new paradigm for fine-tuning reasoning models by replacing the traditional critic network with group-level reward normalization (Guo et al., 2025). While this approach stabilizes updates by computing relative advantages within response sets, its practical application is often hindered by high computational costs and a lack of sensitivity to prompt difficulty. To mitigate these efficiency bottlenecks, CPPO (Lin et al., 2025) introduces completion pruning to discard low-advantage trajectories, while GSPO (Zheng et al., 2025) shifts from token-level to sequence-level importance ratios to stabilize training in Mixture-of-Experts (MoE) architectures. Furthermore, SEED-GRPO (Chen et al., 2025b) integrates semantic entropy into the optimization process, enabling the model to explicitly differentiate between certain and uncertain knowledge boundaries instead of treating all prompts as equally informative.

Building upon these architectural efficiencies, recent research has focused on enhancing the precision of the learning signal through advanced reward shaping and sampling strategies. GRPO-LEAD (Zhang & Zuo, 2025) addresses the issues of verbosity and sparsity by integrating length-regularized rewards and difficulty-aware advantage reweighting, which ensures robust generalization on challenging problems. Similarly, DAPO (Yu et al., 2025) introduces a decoupled and dynamic sampling system designed to stabilize long Chain-of-Thought (CoT) reasoning. By employing asymmetric clipping to prevent entropy collapse and overlong reward shaping, these methods collectively evolve GRPO into a more resilient framework capable of handling complex, multi-step logical synthesis.

While the aforementioned studies focus on algorithmic refinements such as reward shaping or sampling efficiency, they implicitly assume a centralized data architecture where global statistics are readily available. In contrast, our proposed FGRPO framework breaks this assumption by enabling collaborative reasoning training across distributed private datasets.