HIERARCHICAL CREDIT ASSIGNMENT FOR RLVR ON FUSED GROMOV-WASSERSTEIN GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has been shown to improve the reasoning capability of large language models (LLMs) across diverse reasoning tasks. However, group-based RLVR methods, such as GRPO, assign a uniform advantage to all tokens within rollouts of the same outcome. While existing works refine credit assignment of GRPO based on local signals such as token positions or entropy, they often fail to capture the *global semantic novelty* of a reasoning behavior relative to the current policy. In this work, we propose a hierarchical credit assignment approach for group-based RLVR methods, called HARA, which identifies and exploits semantically novel reasoning behaviors during RLVR. HARA represents each sampled rollout as a distribution over the hidden states and positions of tokens, and computes the Fused Gromov-Wasserstein (FGW) barycenters of all rollouts with the same outcome, capturing the internal reasoning patterns in the latent space under the current policy. The semantic novelty of a reasoning element can then be measured by its contribution to the FGW distance between the current rollout and the barycenter. While solving the FGW formulation is expensive, we introduce an anchor-guided linearization that turns it into a Wasserstein formulation solvable via the Sinkhorn algorithm efficiently. By reweighing token-level advantage of group-based RLVR methods based on the novelty signals, HARA exploits novel reasoning behaviors at flexible granularities to encourage fine-grained LLMs exploration. Extensive experiments across three group-based RLVR methods show that our plug-and-play method effectively enhances the exploration of LLMs, outperforming existing methods across diverse reasoning benchmarks.

1 INTRODUCTION

Recent advances in large language models (LLMs) have revealed remarkable potentials of reinforcement learning with verifiable rewards (RLVR) (Schulman et al., 2017; Shao et al., 2024) in empowering LLMs with complex reasoning capabilities. Group-based RLVR methods (Shao et al., 2024; Liu et al., 2025; Yu et al., 2026), such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024), samples multiple rollouts to estimate their relative advantages by grouping them based on binary rewards from a verifier. Such approaches eliminate the need for training an explicit critic model, and has become a major paradigm for RLVR. (Zhang et al., 2025b;a; Wei et al., 2026).

Despite the effectiveness of group-based RLVR methods, they assign uniform advantage to different sampled rollouts and their consisting tokens due to the sparsity of verifiable rewards, providing coarse-grained credit assignment that limits the exploration of LLMs during post-training. While there exists works which aim to better characterize rollout and/or token-level credits to encourage better exploration, they bear the following key limitations: **1) Locality of the signals.** Existing credit assignment methods for GRPO mostly rely on local statistics within individual rollouts, such as token position (Parthasarathi et al., 2025) or entropy (Cheng et al., 2026; Xie et al., 2026), for advantage reshaping. While advantage estimation can propagate such signals across tokens, it remains local within each rollout thus cannot capture the novelty of a reasoning token globally relative to the current policy. **2) Surface-level modeling.** The position and entropy-based uncertainty of a token largely encode the surface-level information of a token, failing to capture its true semantic novelty For example, as shown in Figure 1, position-based methods emphasize prior tokens which hardly impact future reasoning states, while entropy-based methods assign higher credits to tokens with

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.

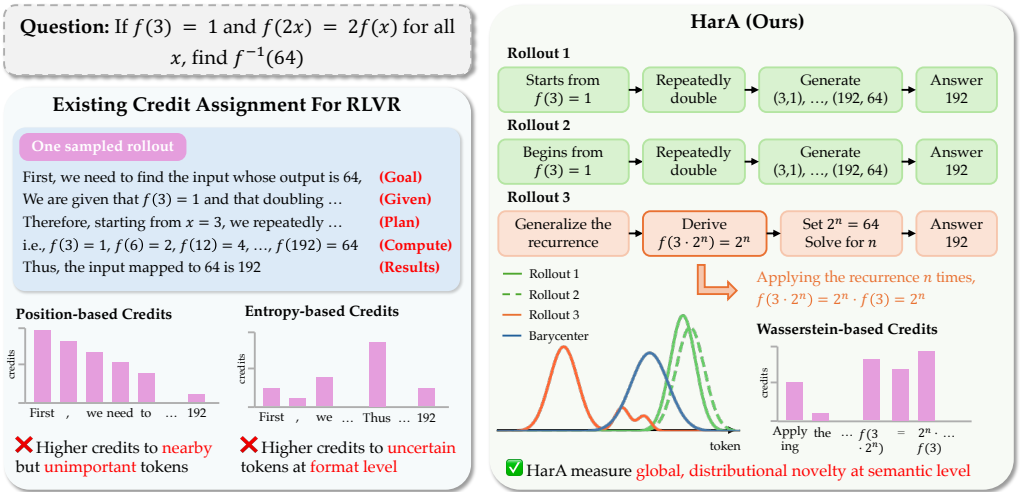


Figure 1: An example comparing existing credit assignment methods and the proposed HARA.

flexible wording choices (e.g.,"Thus") rather than tokens that leads to meaningful exploration. **3) Lack of multi-scale attribution.** Novel reasoning naturally occurs at multiple scales, from individual tokens, steps, to complete reasoning trajectories. However, existing credit assignment methods typically operates at a single granularity, assigning rollout (Liu et al., 2026), span (Chen et al., 2026), or token-level credits (Parthasarathi et al., 2025; Cheng et al., 2026) without a principled mechanism for attribution across multiple granularities.

To tackle the above limitations of existing credit assignment methods, we propose a hierarchical credit assignment approach for group-based RLVR called HARA, leveraging Optimal Transport (OT) as a principled tool for identifying and encouraging reasoning novelty at arbitrary granularity. HARA goes beyond local signals by representing each rollout as a distribution over tokens and comparing it to a Fused Gromov-Wasserstein (FGW) barycenter of all rollouts of the same outcome, capturing the global novelty of a reasoning element relative to the current policy (**Limitation #1**). Moreover, by encoding the transportation cost via both the hidden states and the structural position of a token, HARA summarizes shared reasoning patterns in the latent space via the FGW barycenter, thus capturing reasoning novelty at both semantical (Hao et al., 2024) and structural levels (**Limitation #2**). While the original FGW formualtion is computationally expensive, we leverage an anchor-guided linearization to turns it into a Wasserstein formulation that can be solved efficiently via the Sinkhorn algorithm (Gabriel & Marco, 2019). Finally, the FGW formulation naturally provide fine-grained soft alignment of tokens to abstract reasoning steps represented by barycenter supports, making it easy to measure the novelty of a reasoning element at arbitrary granularity based on its FGW distance to the barycenter (**Limitation #3**). By reweighing token-level advantages based on the novelty measure, HARA serves as a direct plug-and-play framework that is widely applicable to diverse group-based RLVR methods (Shao et al., 2024; Liu et al., 2025; Yu et al., 2026).

Comprehensive experiments across 3 model sizes, 5 reasoning benchmarks, and 3 group-based RLVR methods demonstrates that HARA avoids premature entropy collapse and leads to more extensive exploration, outperforming existing credit assignment approaches while maintaining a good balance between effectiveness and efficiency. Moreover, our observation in Figure 2 shows that the FGW barycenter encodes internal reasoning patterns of the LLM, making them principle references for identifying novel reasoning behaviors that contribute to explorations of LLMs during RLVR. The main contributions of this paper are summarized as follows:

- **Barycenter View of LLMs.** To our best knowledge, we are the first to represent internal reasoning patterns of LLMs via the FGW barycenter of its sampled rollouts in the latent space.
- **Multi-scale Credit Assignment.** We propose a hierarchical credit assignment approach HARA for group-based RLVR methods that idenitfy novel reasoning behaviors at arbitrary granularity.
- **Extensive Analysis.** We provide extensive analysis on the approximation error of the anchor-guided linearization and the complexity of HARA.
- **Strong Performance.** Extensive experiments across diverse model sizes, reasoning benchmarks, and group-based RLVR methods demonstrate that HARA consistently outperforms exitsing credit assignment approaches by a large margin.

2 PRELIMINARIES

2.1 REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS (RLVR)

RLVR has played a critical role in LLMs post-training to empower them with complex reasoning capabilities (Schulman et al., 2017; Shao et al., 2024). Formally, let π_θ denotes the policy of an LLM parameterized by θ. Given a prompt x and its ground-truth answer a, the LLM generates a response sequence $y = (y_1, y_2, ..., y_l)$ of l tokens autoregressively under its policy, i.e., $\pi_\theta(y|x) = \prod_{i=1}^{l} \pi_\theta(y_i|x, y_{<i})$, and is evaluated based on a reward function $R(y, a) = \mathbb{1}\{y \equiv a\}$ where $\mathbb{1}\{\cdot\}$ denotes a binary indictor function of whether the outcome of response y is equivalent to a.

2.2 GROUP-BASED POLICY OPTIMIZATION

Group-based Policy Optimization, a dominant paradigm in RLVR, estimates the relative advantages of multiple sampled rollouts by grouping them based on verifiable rewards without an explicit critic model (Shao et al., 2024; Liu et al., 2025; Yu et al., 2026). One representative method is Group Relative Policy Optimization (GRPO) (Shao et al., 2024), whose objective is defined as

$$\mathbb{E}_{x,\{y_k\}_k^K \sim \pi_\theta(\cdot|x)} \left[\frac{1}{K} \sum_{k=1}^{K} \frac{1}{|y_k|} \sum_{i=1}^{|y_k|} \min\left(r_{k,i} A_k, \text{clip}\left(r_{k,i}, 1-\epsilon, 1+\epsilon\right) A_k\right) \right] - \beta \cdot \text{KL}\left(\pi_\theta(\cdot|x) \| \pi_{\text{ref}}(\cdot|x)\right)$$

where $\{y_k\}_k^K \sim \pi_\theta(\cdot|x)$ denotes K sampled rollouts, $r_{k,i} = \frac{\pi_\theta(y_{k,i}|x, y_{k,<i})}{\pi_{\theta_{\text{old}}}(y_{k,i}|x, y_{k,<i})}$ denotes the importance ratio where $\pi_{\theta_{\text{old}}}$ is the behavior policy from the last policy gradient update, and $A_k = \frac{R(y_k,a) - \text{mean}\left(\{R(y_k,a)\}_{k=1}^K\right)}{\text{std}\left(\{R(y_k,a)\}_{k=1}^K\right)}$ denotes the relative advantage of the k-th sampled rollout. $\text{KL}(\cdot\|\cdot)$ denotes the KL divergence, π_{ref} denotes the reference policy, and β denotes the regularization weight.

2.3 FUSED GROMOV-WASSERSTEIN BARYCENTER

The FGW distance serves as a powerful tool for measuring the discrepancy between geometric data, such as graphs (Titouan et al., 2019; Zeng et al., 2025), by taking into account both node feature and graph structural information. In this work, we represent each LLM rollout as an attributed graph to jointly capture their semantic and structural information. Formally, an attributed graph $\mathcal{G} = \{\mathbf{A}, \mathbf{X}\}$, where $\mathbf{A}$ denotes the graph adjacency matrix and $\mathbf{X}$ denotes the node feature matrix, can be represented by a discrete distribution on the product space of node feature and structure, i.e., $\boldsymbol{\mu} = \sum_{i=1}^{l} \mathbf{w}(i) \delta_{\mathbf{A}(v_i), \mathbf{X}(v_i)}$, where $\mathbf{w} \in \Delta_l$ is a histogram representing the node weight of node $v_i \in \mathcal{G}$, and δ is the Dirac function. In this way, the FGW distance, which combines the Wasserstein (W) and the Gromov-Wasserstein (GW) distance linearly, can be formally defined as follows

Definition 2.1. *Fused Gromov-Wasserstein (FGW) distance*
Given: (1) two attributed graphs $\mathcal{G}_1, \mathcal{G}_2$ with distributions $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, (2) a structural distance measure $f : \mathcal{G}_i \times \mathcal{G}_i \to \mathbb{R}_{\geq 0}$ which captures pairwise node distances within the same graph based on graph structure $\mathbf{A}_i$ for $i \in \{1, 2\}$, e.g., shortest-path distance, (3) a feature distance measure $g : \mathcal{G}_1 \times \mathcal{G}_2 \to \mathbb{R}_{\geq 0}$ which captures pairwise node distances across two graphs based on node features $\mathbf{X}_1$ and $\mathbf{X}_2$, e.g., L_2 distance, (4) the order of FGW distance q, (5) hyperparameter $0 < \alpha < 1$
Output: the FGW distance $\text{FGW}_{q,\alpha}(\mathcal{G}_1, \mathcal{G}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, f, g)$ between $\mathcal{G}_1$ and $\mathcal{G}_2$

$$\text{FGW}_{q,\alpha}(\mathcal{G}_1, \mathcal{G}_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, f, g) = \min_{\mathbf{T} \in \Pi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)} (1-\alpha)\text{W}_q(\mathbf{T}, g) + \alpha \text{GW}_q(\mathbf{T}, f) \tag{1}$$

$$\text{s.t. } \text{W}_q(\mathbf{T}, g) = \mathbb{E}_{(u,v)\sim\mathbf{T}}\left[g^q(u, v)\right], \quad \text{GW}_q(\mathbf{T}, f) = \mathbb{E}_{(u,v),(u',v')\sim\mathbf{T}\otimes\mathbf{T}}\left[\left(f(u, u') - f(v, v')\right)^q\right]$$

where $u, u' \in \mathcal{G}_1$ and $v, v' \in \mathcal{G}_2$ are nodes in the two graphs. We adopt $q = 2$ throughout the paper. Given multiple graphs, the FGW barycenter serves as a consensus graph by summarizing their shared patterns in node features and graph structures based on FGW distance. Formally,

Definition 2.2. *Fused Gromov-Wasserstein (FGW) barycenter* (Zeng et al., 2024)
Given: (1) a set of attributed graphs $\{\mathcal{G}_k = \{\mathbf{A}_k, \mathbf{X}_k\}\}_{k=1}^K$ with distributions $\{\boldsymbol{\mu}_k\}_{k=1}^K$, (2) an intra-graph structural measure f, (3) an cross-graph feature measure g, (4) a barycenter distribution $\boldsymbol{\mu}_b$ with a fixed weight vector $\mathbf{w}_b \in \Delta_{l_b}$, (5) the order of FGW distance q, (6) hyperparameter α
Output: the FGW barycenter $\mathcal{G}_b = \{\mathbf{A}_b, \mathbf{X}_b\}$

$$\{\mathbf{A}_b, \mathbf{X}_b\} := \arg\min_{\mathbf{A},\mathbf{X}} \sum_{k=1}^{K} \text{FGW}_{q,\alpha}\left(\{\mathbf{A}_k, \mathbf{X}_k\}, \{\mathbf{A}, \mathbf{X}\}, \boldsymbol{\mu}_k, \boldsymbol{\mu}_b, f, g\right) \tag{2}$$

3 METHODOLOGY

In this section, we present the proposed credit assignment approach HARA for group-based RLVR methods. Firstly, we introduce a principled way of representing the internal reasoning pattern of LLM within an outcome-grounded group based on the FGW barycenters, in Section 3.1. Then, we introduce a hierarchical credit assignment approach based on the FGW distance to the barycenter, in Section 3.2. Finally, we provide theoretical analysis regarding the approximation error and the complexity of HARA, in Section 3.3.

3.1 DISTRIBUTIONAL MODELING OF INTERNAL REASONING PATTERN

Recent studies (Hao et al., 2024; Du et al., 2026) show that the hidden states of an LLM encode abstract reasoning steps in a continuous latent space and can signal when reasoning behavior diverges. Such observations imply that novel reasoning behaviors may be discovered by comparing the hidden states of sampled rollouts in the latent space. In this work, we propose to leverage OT as a principled tool for such comparisons. Specifically, we leverage the FGW barycenters of sampled rollouts within an outcome-grounded group to encode the internal reasoning pattern of the current policy, which later serves as a principled reference for measuring the novelty of a reasoning element.

3.1.1 FGW BARYCENTER AS INTERNAL REASONING PATTERN

LLM rollouts are in the form of natural language, making principled comparison across different rollouts challenging and costly. To address this, we project each rollout into a distribution on the product space of the hidden states and position of its tokens, jointly capturing the semantic and structural information of a rollout in the latent space. Formally, let y denotes a rollout of l tokens, we represent y as an attributed chain graph $\mathcal{G} = \{\mathbf{A}, \mathbf{X}\}$ with l nodes, where $\mathbf{A}[i,j] = \mathbb{1}\{|i-j| = 1\}$ represents its chain structure, and the node feature $\mathbf{X} = \mathbf{H} \in \mathbb{R}^{l \times d}$ is the hidden state matrix of y.

After representing rollouts as distributions, we leverage FGW barycenters to summarize the internal reasoning patterns of the current policy within an outcome-grounded group. Formally, for a prompt x with ground-truth answer a, let $\{\mathcal{G}_k^{(c)} = \{\mathbf{A}_k^{(c)}, \mathbf{X}_k^{(c)}\}\}_{k=1}^K$ denotes the graph representations of K sampled rollouts $\{y_k^{(c)} | R(y_k^{(c)}, a) = \mathbb{1}\{c = +\}\}_{k=1}^K$ within group $c \in \{+, -\}$, i.e., sets of correct and incorrect rollouts, we compute a FGW barycenter $\mathcal{G}_b^{(c)} = \{\mathbf{A}_b^{(c)}, \mathbf{X}_b^{(c)}\}$ for group c as follows

$$\mathbf{X}_b^{(c)} := \arg\min_{\mathbf{X}} \sum_{k=1}^{K} \mathrm{FGW}_{2,\alpha}\left(\{\mathbf{A}_k^{(c)}, \mathbf{X}_k^{(c)}\}, \{\mathbf{A}_b^{(c)}, \mathbf{X}\}, f, g\right) \tag{3}$$

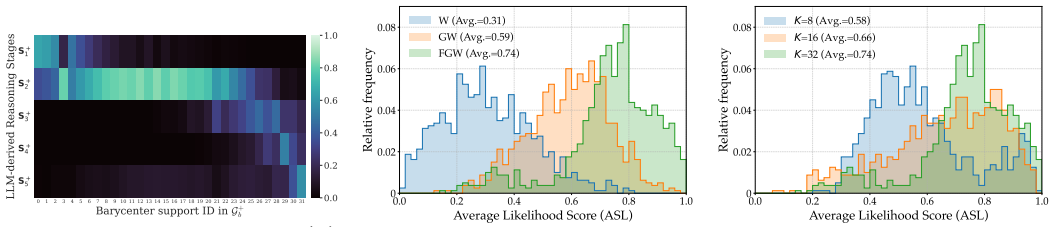
where $f(y_{k,i}^{(c)}, y_{k,j}^{(c)}) = |i-j|/(l_k^{(c)} - 1)$ defines a normalized shortest-path distance of tokens within each rollout, and $g(y_{k,i}^{(c)}, y_{b,j}^{(c)}) = \|\mathbf{X}_k^{(c)}[i,:] - \mathbf{X}_b^{(c)}[j,:]\|_2$ defines a L_2 distance between hidden states of tokens in sampled rollouts and the barycenter. Note that compared to Eq. (2), we fix $\mathbf{A}_b^{(c)}[i,j] = \mathbb{1}\{|i-j| = 1\}$ to a chain structure to represent a *mean rollout* in the latent space, and only optimize the hidden states matrix $\mathbf{X}_b^{(c)}$ of the barycenter in Eq. (3).

However, the FGW barycenter is inherently a discrete distribution that lies in a latent space, making it less interpretable than an actual rollout consisting of natural language tokens. To gain a deeper insight into what the FGW barycenter represents, we conduct the following interpretability studies to examine whether its individual supports correspond to meaningful reasoning stages in natural language. Specifically, we obtain an OT transport plan $\mathbf{T}_k^{(c)}$ for $y_k^{(c)}$ when solving Eq. (3), which serves as a soft mapping between tokens in $y_k^{(c)}$ and the barycenter. To further establish a connection between barycenter supports and meaningful reasoning stages, we prompt a powerful LLM, i.e., GPT-5-Sol, to derive a set of shared reasoning stages $\{S_m^{(c)}\}_{m=1}^M$ from all sampled rollouts in group c in natural languages, and classify each token to one reasoning stages, e.g., $y_{k,i}^{(c)} \in S_m^{(c)}$. In this way, we obtain a mapping $\mathbf{P}^{(c)}$ between barycenter supports and LLM-derived reasoning stages:

$$\mathbf{P}^{(c)}(j,m) = \frac{\sum_{k=1}^{K} \sum_{i=1}^{|y_k^{(c)}|} \mathbf{T}_k^{(c)}[i,j] \mathbb{1}\left\{y_{k,i}^{(c)} \in S_m^{(c)}\right\}}{\sum_{k=1}^{K} \sum_{i=1}^{|y_k^{(c)}|} \mathbf{T}_k^{(c)}[i,j]} \tag{4}$$

We visualize $\mathbf{P}^{(+)}$ for one question [1] in MATH500 (Hendrycks et al., 2021) in Figure 2a, which shows a *sparse* and *block-diagonal* matrix indicating that the FGW barycenter supports can be

[1] **Question**: If $f(3) = 1$ and $f(2x) = 2f(x)$ for all x, find $f^{-1}(64)$



(a) Visualization of mapping $\mathbf{P}^{(+)}$ (b) Comparison of OT variants. (c) Comparison of sample size K.

Figure 2: Observations. (1) **Visualization of mapping $\mathbf{P}^{(+)}$**: The *sparse* mapping shows that the barycenter supports are mapped deterministically to meaningful reasoning stages. (2) **Comparison of OT variants**: FGW exhibits higher ASL than W and GW, demonstrating stronger expressiveness. (3) **Comparison of sample size** K: ASL improves monotonically as K increases, suggesting that the barycenter reliably captures the internal reasoning pattern with sufficient sampling.

mapped deterministically to meaningful reasoning stages. To further evaluate the accuracy of such mapping, we design a Leave-One-Rollout-Out (LORO) evaluation where we randomly hold out one sampled rollout $y_e^{(c)}$ when constructing the mapping $\mathbf{P}^{(c)}$, and then evaluate $\mathbf{P}^{(c)}$ by its average likelihood score (ASL) in predicting token labels in $y_e^{(c)}$ as follows:

$$\mathrm{ASL}(y_e^{(c)}) = \sum_{i=1}^{|y_e^{(c)}|} \sum_{j=1}^{l_b} \sum_{m=1}^{M} \mathbf{T}^{(c)}[i,j]\mathbf{P}^{(c)}[j,m]\mathbb{1}\{y_{e,i}^{(c)} \in S_m^{(c)}\} \tag{5}$$

The comparative results of LORO evaluation across different OT variants and sampling sizes on MATH500 are included in Figure 2b and Figure 2c, respectively. We observe that **1)** The FGW formulation achieve higher ASL than the Wasserstein (W) and Gromov-Wasserstein (GW) formulation, demonstrating stronger expressiveness by combining both semantic hidden states and structural positions of tokens; **2)** The ASL under FGW formulation improves monotonically as K increases, suggesting that the barycenter reliably captures the internal reasoning pattern of current policy with sufficient sampling. Detailed experimental settings of Figure 2 are included in Appendix B.4.

3.1.2 Anchor-Guided Linearization

The FGW formulation in Eq. (1), while expressive, is computationally expensive due to its quartic, non-convex GW term (Peyré et al., 2016) of $O(l^3)$ time complexity to solve. To optimize efficiency, we leverage an anchor-guided linearization technique (Im et al., 2026) that linearizes the GW term by fixing its second transport plan to pre-align reasoning state anchors. Specifically, we treat the first and last tokens of each sampled rollout, i.e., $y_{k,1}^{(c)}$ and $y_{k,L_k}^{(c)}$, as reasoning state anchors, as they represent the shared initial and final states of all sampled rollouts under the same prompt x and outcome c. By pre-aligning anchors in the GW term in Eq. (1), we can linearize the FGW distance in Eq. (3) to a Wasserstein distance as follows:

Lemma 3.1 (Anchor-Guided Linearization of FGW Distance)**.** $\mathrm{FGW}_{2,\alpha}$ [2] *in Eq.* (3) *can be linearized into the following Wasserstein distance* $\mathrm{FGW}_{2,\alpha}^{linear}$ *by fixing the second transport plan* $\mathbf{T}_{anchor}^{(c)} = \frac{1}{2}\left(\delta_{(1,1)} + \delta_{(L_k,L_b)}\right)$ *in its* GW_2 *term that pre-aligns reasoning anchors* $y_{k,1}^{(c)}$ *and* $y_{k,l_k}^{(c)}$.

$$\mathrm{FGW}_{2,\alpha}^{\mathrm{linear}} = \min_{\mathbf{T}\in\Pi(\boldsymbol{\mu}_k,\boldsymbol{\mu}_b)} \mathbb{E}_{\left(y_{k,i}^{(c)},y_{b,j}^{(c)}\right)\sim\mathbf{T}}\left[(1-\alpha)g^2\left(y_{k,i}^{(c)},y_{b,j}^{(c)}\right) + 2\alpha\left|\frac{i-1}{L_k-1} - \frac{j-1}{L_b-1}\right|^2\right] \tag{6}$$

Detailed proof of Lemma 3.1 can be found in Appendix A.1. In general, fixing the second transport plan in GW_2 turns it into a convex Wasserstein term that penalizes the difference in normalized positions of tokens across different rollouts, which can be solved efficiently via an entropy regularization and the Sinkhorn algorithm (Gabriel & Marco, 2019). In this way, we can obtain the linearized FGW barycenter efficiently by replacing $\mathrm{FGW}_{2,\alpha}$ in Eq. (3) with $\mathrm{FGW}_{2,\alpha}^{\mathrm{linear}}$ in Eq. (6).

3.2 Hierarchical Credit Assignment

After obtaining the linearized FGW barycenters as principled reference distributions, we measure the novelty of a reasoning span at arbitrary granularity by their contributions to the linearized

[2]For brevity, we use $\mathrm{FGW}_{2,\alpha}$ as an abbreviation for $\mathrm{FGW}_{2,\alpha}(\{\mathbf{A}_k^{(c)},\mathbf{X}_k^{(c)}\},\{\mathbf{A}_b^{(c)},\mathbf{X}_b^{(c)}\},f,g)$

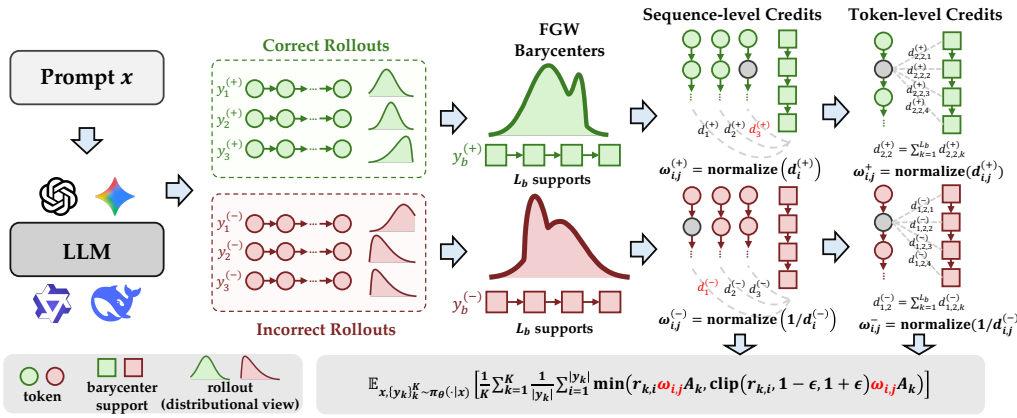


Figure 3: Overview of HARA. FGW barycenters are computed from the correct and incorrect rollouts as two reference distributions, followed by a hierarchical credit assignment approach that reweights the group advantages at both the sequence and token levels, based on their reasoning novelty measured by the linearized FGW distance to the barycenters.

FGW distance to the barycenters, and propose a hierarchical credit assignment approach for group-based RLVR methods as shown in Figure 3. Formally, let $\mathcal{P}_k^{(c)}$ be a partition of rollout $y_k^{(c)}$ into $M_k^{(c)} = |\mathcal{P}_k^{(c)}|$ reasoning spans, the novelty-based credit $\omega_{k,\mathcal{S}}^{(c)}$ for a span $y_{k,\mathcal{S}}^{(c)} \subseteq y_k^{(c)}$ is defined as its contribution to the linearized FGW distance to $\mathcal{G}_b^{(c)}$ normalized within group c as follows

$$\omega_{k,\mathcal{S}}^{(c)} = \frac{\left(\sum_{k'=1}^{K} M_{k'}^{(c)}\right)\left(\sum_{i\in\mathcal{S}} \phi_{k,i}^{(c)}\right)^p}{\sum_{k'=1}^{K}\sum_{\mathcal{S}'\in\mathcal{P}_k^{(c)}}\left(\sum_{i'\in\mathcal{S}'} \phi_{k',i'}^{(c)}\right)^p},$$
$$\text{where} \quad \phi_{k,i}^{(c)} = \sum_{j=1}^{L_b} \mathbf{T}_k^{(c)}(i,j)\left[(1-\alpha)\left\|\mathbf{X}_{k,i}^{(c)} - \mathbf{X}_{b,j}^{(c),*}\right\|_2^2 + 2\alpha\left(\frac{i-1}{L_k-1} - \frac{j-1}{L_b-1}\right)^2\right]. \tag{7}$$

where $p = 1$ for correct rollouts (i.e., $c = +$) and $p = -1$ for incorrect rollouts (i.e., $c = -$). The intuition of p is to guide the RLVR process to encourage correct and novel reasoning behaviors more, while penalizing incorrect but novel reasoning behaviors less to preserve the exploration of the LLM. Based on the obtained novelty-based credits $\omega_{k,\mathcal{S}}^{(c)}$, we boadcast the same credit to all tokens inside the reasoning span $\mathcal{S}$, i.e., $\omega_{k,i}^{(c)} = \omega_{k,\mathcal{S}}^{(c)}$ for all $i \in \mathcal{S}$, and reweight the token-level group advantages for group-based RLVR methods as follows

$$\mathbb{E}_{x,\{y_k\}_k^K\sim\pi_\theta(\cdot|x)}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|y_k|}\sum_{i=1}^{|y_k|}\min\left(r_{k,i}\omega_{k,i}A_k, \text{clip}\left(r_{k,i}, 1-\epsilon, 1+\epsilon\right)\omega_{k,i}A_k\right)\right] \tag{8}$$

In this paper, we mainly consider credit assignment on two levels of granularity: *sequence-level* (i.e., $\mathcal{P}_k^{(c)} = \{y_k^{(c)}\}$) and *token-level* (i.e., $\mathcal{P}_k^{(c)} = \{y_{k,i}^{(c)}\}_{i=1}^{l_k}$), and name the corresponding variants as HARA-seq and HARA-tok, respectively.

3.3 THEORETICAL ANALYSIS

In this subsection, we provide theoretical analysis regarding the approximation error of anchor-guided linearization and the complexity of the proposed HARA.

Theorem 3.2 (Approximation Error of Anchor-Guided Linearization). *Let $\mathbf{T}_{\text{FGW}}^* \in \Pi(\mu_k, \mu_b)$ be an optimal transport plan that minimizes the original FGW problem with uniform marginal distributions μ_k and μ_b. The approximation error of the anchor-guided linearization in Lemma 3.1 is bounded as follows*

$$0 \le \text{FGW}_{2,\alpha}^{linear}(\mathcal{G}_k^{(c)}, \mathcal{G}_b^{(c)}) - \text{FGW}_{2,\alpha}(\mathcal{G}_k^{(c)}, \mathcal{G}_b^{(c)}) \le 4\alpha\,\mathcal{C}(\mathbf{T}_{\text{FGW}}^*), \tag{9}$$

where

$$\mathcal{C}(\mathbf{T}) = \mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[\frac{|I-I'|}{L_k-1}\frac{|J-J'|}{L_b-1}\mathbb{1}\{(I-I')(J-J') < 0\}\right], \qquad (I,J),(I',J') \overset{\text{i.i.d.}}{\sim} \mathbf{T}. \tag{10}$$

The detailed proof of Theorem 3.2 can be found in Appendix A.2. In general, the upper-bound of the approximation error is determined by the magnitude of *order-reversing entries* in $\mathbf{T}^*_{\mathrm{FGW}}$, i.e., pairs of (I, J) and (I', J') such that $(I - I')(J - J') < 0$. Therefore, the error is expected to remain small as the FGW term in Eq. (3) encourages alignment between tokens of similar normalized positions, thus producing a *sparse* and *block-diagonal* transport plan $\mathbf{T}^*_{\mathrm{FGW}}$ with few order-reversing entries.

Theorem 3.3 (Complexity of HARA)**.** *The time complexity of* HARA *is* $\mathcal{O}(BKLL_b(d + S))$*, and the space complexity of* HARA *is* $\mathcal{O}(KLL_b + KLd + L_bd)$*, where* K *is the number of sampled rollouts of length* L*,* L_b *is the barycenter size,* d *is the hidden state dimension.* S *and* B *are the number of Sinkhorn and barycenter optimization iterations for solving Eq.* (6)*, respectively.* [3].

The detailed proof of Theorem 3.3 can be found in Appendix A.3. Notably, HARA scales *linearly* w.r.t. the rollout length L in both time and space. With a fixed and relatively small barycenter size $L_b = 32$, HARA enables efficient and scalable credit assignment for group-based RLVR methods across various reasoning tasks, as shown in Section 4.3.

4 EXPERIMENTS

In this section, we carry out extensive experiments and analysis to evaluate the proposed HARA from the following aspects:

- **Q1:** How effective is HARA compared to existing credit assignment approaches?
- **Q2:** How does HARA affect the training dynamics of group-based RLVR methods?
- **Q3:** How does HARA balance between effectiveness and efficiency?
- **Q4:** How good is the FGW distance as a novelty metrics for HARA?
- **Q5:** How sensitive is HARA to different hyperparameter settings and design choices?

4.1 EXPERIMENTAL SETUP

For **datasets**, we train group-based RLVR methods on the MATH dataset (Hendrycks et al., 2021), and evaluate HARA across five different math reasoning benchmarks, including MATH500, AMC23, AIME24, AIME25 (Lightman et al., 2024), and Minerva Math (Lewkowycz et al., 2022); for **models**, we benchmark across three models, including Qwen3-1.7B-Base, Qwen3-4B-Base, and Qwen3-8B-Base (Yang et al., 2025); for **baselines**, we compare HARA with three existing credit assignment methods, including GRPO-λ (Parthasarathi et al., 2025), Entropy Adv. (Cheng et al., 2026), and SHEAR (Chen et al., 2026), when applied to three different group-based RLVR methods, including GRPO (Shao et al., 2024), Dr.GRPO (Liu et al., 2025), and DAPO (Yu et al., 2026). GRPO is adopted as the default RLVR baseline in all experiments unless noted otherwise; for **evaluation metrics**, we adopt the average accuracy for MATH500 and Minerva Math, and the Pass@16 for AMC23, AIME24, and AIME25. More detailed experimental setup are provided in Appendix B.

4.2 MAIN RESULTS

Benchmark Results. The benchmark results of HARA and existing credit assignment methods on GRPO (Shao et al., 2024) are shown in Table 1. We observe that **1)** HARA **consistently achieves the best performance across all datasets and LLMs.** Specifically, HARA consistently improves GRPO by up to 4.8% in the average performance, while also outperforming all credit assignment baselines with up to 4.0% improvement. **2) Local token signals alone can be misleading for credit assignment.** While all credit assignment baselines, on average, improves the original GRPO, they can sometimes degrade the performance of GRPO in individual cases, e.g., GRPO-λ on Qwen3-4B-Base for AIME24, and Entropy Adv. on Qwen3-1.7B-Base for MATH500, showing that relying solely on local structural or uncertainty measure of tokens for credit assignment may be insufficient or even harmful. **3) Effective credit assignment at token level leads to superior performance.** We can see that HARA-tok outperforms HARA-seq on most datasets and LLMs, demonstrating that our method effectively captures novel reasoning behaviors at token level, thus leading to improved reasoning performance. Additional results on Dr.GRPO and DAPO are provided in Appendix C.1, which further demonstrates the effectiveness of HARA.

[3]Without loss of generality, we assume $\mathcal{O}(K^{(+)}) \approx \mathcal{O}(K^{(-)}) \approx \mathcal{O}(K)$ and $\mathcal{O}(L_k^{(c)}) \approx \mathcal{O}(L)$ for all $k \in [1, K]$ and $c \in \{+, -\}$.

Table 1: Benchmark results of credit assignment methods on GRPO (Shao et al., 2024). The 1st/2nd best results are highlighted in **bold**/underline, respectively. Our methods are highlighted in red box.

Method	MATH500	Minerva Math	AMC23	AIME24	AIME25	Average
Qwen3-1.7B-Base						
GRPO	64.8 ±2.8	18.2 ±1.7	72.5 ±2.2	14.0 ±1.3	12.7 ±2.4	36.4
+ GRPO-λ	65.7 ±1.2	17.6 ±1.5	72.0 ±1.0	16.0 ±0.6	14.0 ±1.2	37.1
+ Entropy Adv.	66.5 ±1.8	19.1 ±1.4	74.0 ±1.8	15.3 ±1.2	13.3 ±2.1	37.6
+ SHEAR	65.9 ±1.3	18.7 ±0.9	74.5 ±0.4	12.7 ±0.9	13.3 ±1.5	37.0
+ HARA-seq	**67.3** ±2.0	20.3 ±1.2	76.0 ±1.4	**18.7** ±0.6	16.0 ±1.3	39.7
+ HARA-tok	67.1 ±2.1	**21.7** ±1.3	**77.0** ±1.6	**18.7** ±0.6	**16.7** ±1.2	**40.2**
Qwen3-4B-Base						
GRPO	78.1 ±1.5	31.9 ±1.8	84.5 ±1.7	33.3 ±1.9	29.3 ±2.1	51.4
+ GRPO-λ	76.9 ±0.3	32.1 ±0.6	86.5 ±1.1	32.7 ±0.6	30.0 ±0.6	51.6
+ Entropy Adv.	77.5 ±1.6	33.0 ±1.1	86.0 ±1.9	34.7 ±1.3	31.3 ±1.3	52.5
+ SHEAR	78.5 ±0.8	32.8 ±1.3	85.0 ±0.9	33.3 ±1.5	30.7 ±1.2	52.1
+ HARA-seq	80.9 ±1.7	**34.7** ±2.3	87.5 ±1.1	36.7 ±1.5	32.7 ±0.9	54.5
+ HARA-tok	**82.1** ±1.1	34.1 ±1.5	**88.0** ±1.3	**37.3** ±1.2	**34.0** ±1.5	**55.1**
Qwen3-8B-Base						
GRPO	81.8 ±1.3	32.2 ±0.9	89.5 ±2.0	52.0 ±1.3	40.0 ±1.7	59.1
+ GRPO-λ	82.1 ±0.1	32.4 ±0.5	90.5 ±0.4	53.3 ±0.6	41.3 ±0.6	59.9
+ Entropy Adv.	83.9 ±0.4	33.8 ±1.3	91.0 ±1.9	52.7 ±1.7	42.7 ±1.3	60.8
+ SHEAR	82.9 ±0.7	32.0 ±1.9	90.0 ±2.0	54.7 ±1.3	42.0 ±1.2	60.3
+ HARA-seq	84.7 ±0.3	34.6 ±1.1	92.5 ±2.4	58.0 ±1.2	43.3 ±1.9	62.6
+ HARA-tok	**85.2** ±0.5	**35.0** ±1.6	**94.0** ±1.9	**58.7** ±1.2	**46.7** ±1.7	**63.9**

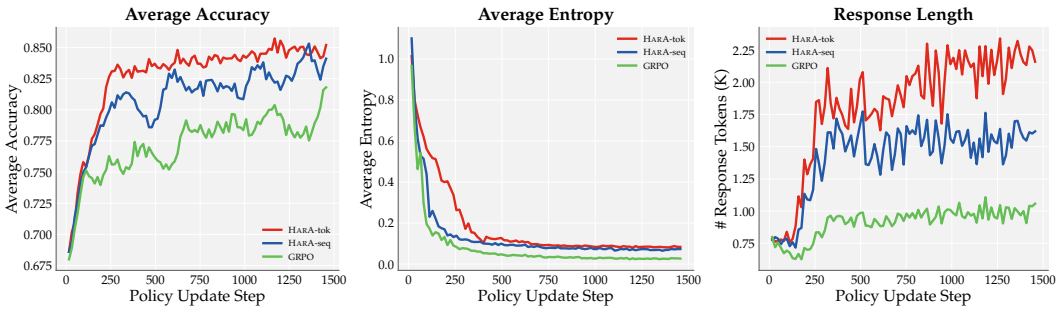


Figure 4: Training dynamics of HARA and GRPO on MATH500 using Qwen3-8B-Base.

Training Dynamics. To further understand how HARA affects the training of group-based RLVR methods, we visualize the training dynamics of GRPO and HARA on MATH500 using Qwen3-8B-Base in Figure 4, tracking three key metrics: *accuracy*, *response length*, and *entropy* of the trained LLM. We can oberve that **1)** HARA **enables faster convergence and improved final performance.** HARA identifies novel reasoning rollouts and tokens to emphasize during training, thus converges faster and achieve sa better final performance than baseline GRPO. **2)** HARA **leads to more extensive exploration.** We can see that HARA maintains higher entropy than baseline GRPO during training, thus preserving more exploration capability of the trained model. Moreover, HARA generates longer responses than baseline GRPO, indicating that it encourages the model to explore more diverse reasoning paths by identifying and emphasizing on novel reasoning behaviors.

4.3 EFFICIENCY RESULTS

As shown in Figure 5, HARA introduces a mild additional overhead of up to 13.6% in total training time compared to baseline GRPO, making it practical to be applied to standard RLVR training. Although HARA does introduce additional overhead, Figure 6 shows that, compared with GRPO, HARA reaches the **same accuracy with up to 2.64× speed-up** in training time, and achieves an **accuracy improvement of 8.7% using the same training time**. This indicates that HARA maintains a good balance between effectiveness and efficiency for improving group-based RLVR methods.

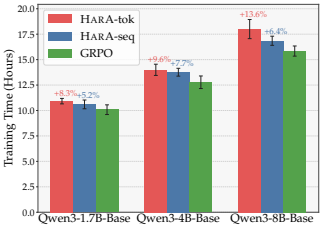


Figure 5: Training overhead of HARA on MATH.

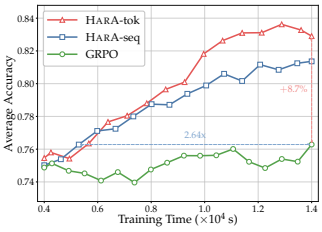


Figure 6: Average accuracy vs. training time on MATH.

Table 2: Comparison of variants of HARA on Qwen3-8B-Base.

Method	MATH500	AMC23
GRPO	81.8	89.5
+ HARA(W)-seq	82.3	90.0
+ HARA(GW)-seq	80.6	87.0
+ HARA(H)-seq	81.2	86.5
+ HARA-seq	**84.7**	**92.5**
+ HARA(W)-tok	83.0	90.5
+ HARA(GW)-tok	79.7	87.0
+ HARA-tok	**85.2**	**94.0**

4.4 FURTHER STUDIES

Comparision of Novelty Metrics. To verify the effectiveness of the FGW formulation in HARA, we compare the performance of different variants of HARA using different novelty metrics for ω in Eq. (8), including Wasserstein (W) and Gromov-Wasserstein (GW) distances in Eq. 1, and an additional metrics defined as the embedding distance between the barycenter and the last hidden state of each rollout (H), i.e., $\omega_{k,i} = \|\mathbf{X}^{(c)}_{k,L_k} - \mathbf{X}^{(c)}_{b'}\|_2^2$ where $\mathbf{X}^{(c)}_{b'} = \frac{1}{K}\sum_{k'=1}^{K} \mathbf{X}^{(c)}_{k',L_k}$. The results are included in Table 2, showing that **1) the FGW distance adopted by HARA is the most effective novelty metric for credit assignment** with consistent outperformance over all other metrics. **2) The GW distance, while effective in aligning tokens in Figure 2, cannot capture semantic novelty of a reasoning behavior**, resulting in degraded performance even compared to baseline GRPO.

Additional Studies. Due to space constraints, we present additional studies of HARA, including **parameters studies** on the FGW trade-off parameter α, and an **ablation study on the anchor-guided linearization**, in Appendix C.2.

5 RELATED WORK

Fused Gromov-Wasserstein Barycenter. Fused Gromov-Wasserstein (FGW) barycenter has been predominantly used in graph learning and time-series modeling. Specifically, (Brogat-Motte et al., 2022) leverage an input-dependent FGW barycenters for interpolatation within the space of attributed graphs; (Cheng et al., 2021) use dynamically weighted Wasserstein barycenters to model gradual transitions in multivariate time series. (Thual et al., 2022) introduce fused unbalanced Gromov-Wasserstein transport to align individual cortical surfaces; (Zeng et al., 2024) leverage an FGW barycenter for cluster-level multi-network alignment.

Credit Assignment for RLVR. Existing credit assignment methods for RLVR operate on sequence, span, and token levels. For sequence-level methods, (Liu et al., 2026) measures intra-group error diversity to reshape the advantage signal of incorrect rollouts. For span-level methods, (Chen et al., 2026) emphasizes on spans that are more distant from the opposing group based on Wasserstein distance, For token-level methods, (Parthasarathi et al., 2025) assigns higher credits to nearby tokens and propagates such credits to preceding tokens with exponential decay, while (Cheng et al., 2026) and (Xie et al., 2026) emphasize on tokens with higher entropy-based uncertainty. While promising, none of them identifies semantically novel reasoning behaviors at the finest token level.

6 CONCLUSION

In this paper, we propose HARA, a novel hierarchical credit assignment method for group-based RLVR methods. HARA represents each sampled rollout as a distribution over the hidden states and locations of tokens, and leverages the Fused Gromov-Wasserstein (FGW) barycenters of all rollouts of the same outcome to capture the internal reasoning patterns in the latent space. By measuring the semantic novelty of a reasoning element based on its contribution to the FGW distance to the barycenter, HARA reweighs token-level advantage of group-based RLVR methods to encourage novel reasoning behaviors at flexible granularities. Extensive experiments demonstrate that our plug-and-play method significantly improves the performance of diverse group-based RLVR methods while maintaining a good balance between effectiveness and efficiency.

AI USE STATEMENT

In this work, generative AI tools are used solely for polishing the grammar and readability of our work. We have not used generative AI tools for other tasks with required or recommended disclosure. We have reviewed all AI-assisted work, and take full responsibility for the final content of this work, including text, claims and artifacts.

REFERENCES

Luc Brogat-Motte, Rémi Flamary, Céline Brouard, Juho Rousu, and Florence d'Alché Buc. Learning to predict graphs with fused gromov-wasserstein barycenters. In *International Conference on Machine Learning*, pp. 2321–2335. PMLR, 2022.

Xinzhu Chen, Wei He, Huichuan Fan, Wenzhe Niu, Zhongxiang Sun, Xuanru Wang, Jiuchong Gao, Jinghua Hao, Renqing He, and Weijie Yu. Hidden states know where reasoning diverges: Credit assignment via span-level wasserstein distance. *arXiv preprint arXiv:2604.23318*, 2026.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30377–30385, 2026.

Kevin Cheng, Shuchin Aeron, Michael C Hughes, and Eric L Miller. Dynamical wasserstein barycenters for time-series modeling. *Advances in Neural Information Processing Systems*, 34: 27991–28003, 2021.

Hanwen Du, Yuxin Dong, and Xia Ning. Latent thinking optimization: Your latent reasoning language model secretly encodes reward signals in its latent thoughts. In *International Conference on Learning Representations*, volume 2026, pp. 101000–101032, 2026.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jiin Im, Sisung Liu, and Je Hyeong Hong. Shape-of-you: Fused gromov-wasserstein optimal transport for semantic correspondence in-the-wild. *arXiv preprint arXiv:2603.11618*, 2026.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Wenpu Liu, Yuqi Xu, Weichu Xie, Yongfu Zhu, Shuai Dong, Ziyue Wang, Wenqi Shao, Xiaoying Zhang, Tong Yang, Nan Duan, et al. Leveraging error diversity in group rollouts for reinforcement learning. *arXiv preprint arXiv:2605.17333*, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. Grpo-λ: Credit assignment improves llm reasoning. *arXiv preprint arXiv:2510.00194*, 2025.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

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.

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in neural information processing systems*, 35:21792–21804, 2022.

Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Dongqi Fu, et al. Agentic reasoning for large language models. *arXiv preprint arXiv:2601.12538*, 2026.

Can Xie, Ruotong Pan, Xiangyu Wu, Zhang Yunfei, Jiayi Fu, Tingting Gao, and Guorui Zhou. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19057–19076, 2026.

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

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. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Zhichen Zeng, Boxin Du, Si Zhang, Yinglong Xia, Zhining Liu, and Hanghang Tong. Hierarchical multi-marginal optimal transport for network alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16660–16668, 2024.

Zhichen Zeng, Ruizhong Qiu, Wenxuan Bao, Tianxin Wei, Xiao Lin, Yuchen Yan, Tarek F Abdelzaher, Jiawei Han, and Hanghang Tong. Pave your own path: Graph gradual domain adaptation on fused gromov-wasserstein geodesics. *arXiv preprint arXiv:2505.12709*, 2025.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025a.

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*, 2025b.

A PROOF

A.1 PROOF OF LEMMA 3.1

Lemma (Anchor-Guided Linearization of FGW Distance)**.** $\mathrm{FGW}_{2,\alpha}$ *in Eq.* (3) *can be linearized into the following Wasserstein distance* $\mathrm{FGW}_{2,\alpha}^{linear}$ *by fixing the second transport plan* $\mathbf{T}_{anchor}^{(c)} = \frac{1}{2}\left(\delta_{(1,1)} + \delta_{(l_k,l_b)}\right)$ *in its* GW_2 *term that pre-aligns reasoning anchors* $y_{k,1}^{(c)}$ *and* $y_{k,l_k}^{(c)}$

$$\mathrm{FGW}_{2,\alpha}^{\mathrm{linear}} = \min_{\mathbf{T}\in\Pi(\boldsymbol{\mu}_k,\boldsymbol{\mu}_b)} \mathbb{E}_{\left(y_{k,i}^{(c)},y_{b,j}^{(c)}\right)\sim\mathbf{T}}\left[(1-\alpha)g^2\left(y_{k,i}^{(c)},y_{b,j}^{(c)}\right) + 2\alpha\left|\frac{i-1}{l_k-1} - \frac{j-1}{l_b-1}\right|^2\right] \tag{11}$$

Proof. Based on Eq. (1), the $\mathrm{FGW}_{2,\alpha}$ objective between the graph representation rollout $\mathcal{G}_k^{(c)}$ of rollout $y_k^{(c)}$ and the barycenter $\mathcal{G}_b$ can be expanded as follows

$$\begin{aligned}\mathrm{FGW}_{2,\alpha} = \min_{\mathbf{T}\in\Pi(\mu_k,\mu_b)} \Bigg\{ &(1-\alpha)\mathbb{E}_{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}}\left[g^2\left(y_{k,i}^{(c)},y_{b,j}^{(c)}\right)\right] \\ &+ \alpha \underbrace{\mathbb{E}_{\substack{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}\\(y_{k,i'}^{(c)},y_{b,j'}^{(c)})\sim\mathbf{T}}}\left[\left|f\left(y_{k,i}^{(c)},y_{k,i'}^{(c)}\right) - f\left(y_{b,j}^{(c)},y_{b,j'}^{(c)}\right)\right|^2\right]}_{\mathrm{GW}_2} \Bigg\}.\end{aligned} \tag{12}$$

$\mathrm{FGW}_{2,\alpha}$ is difficult to solve as GW_2 is non-convex and quadratic w.r.t $\mathbf{T}$. To improve efficiency, we linearize GW_2 by fixing its second transport plan to $\mathbf{T}_{\mathrm{anchor}}^{(c)} = \frac{1}{2}\left(\delta_{(1,1)} + \delta_{(l_k,l_b)}\right)$ that pre-aligns the reasoning anchors by assigning mass only to the two entries $(1,1)$ and (l_k,l_b) that aligns $(y_{k,1}^{(c)}, y_{b,1}^{(c)})$ and $(y_{k,l_k}^{(c)}, y_{b,l_b}^{(c)})$, In this way, the linearized GW term $\mathrm{GW}_2^{\mathrm{linear}}$ can be derived as follows

$$\begin{aligned}\mathrm{GW}_2^{\mathrm{linear}} &= 2\cdot\mathbb{E}_{\substack{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}\\(y_{k,i'}^{(c)},y_{b,j'}^{(c)})\sim\mathbf{T}_{\mathrm{anchor}}^{(c)}}}\left[\left|f\left(y_{k,i}^{(c)},y_{k,i'}^{(c)}\right) - f\left(y_{b,j}^{(c)},y_{b,j'}^{(c)}\right)\right|^2\right] \\ &= \mathbb{E}_{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}}\left[\sum_{(i',j')\in\{(1,1),(l_k,l_b)\}}\left|f\left(y_{k,i}^{(c)},y_{k,i'}^{(c)}\right) - f\left(y_{b,j}^{(c)},y_{b,j'}^{(c)}\right)\right|^2\right] \\ &= \mathbb{E}_{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}}\left[\left|f\left(y_{k,i}^{(c)},y_{k,1}^{(c)}\right) - f\left(y_{b,j}^{(c)},y_{b,1}^{(c)}\right)\right|^2 + \left|f\left(y_{k,i}^{(c)},y_{k,l_k}^{(c)}\right) - f\left(y_{b,j}^{(c)},y_{b,l_b}^{(c)}\right)\right|^2\right] \\ &= \mathbb{E}_{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}}\left[\left|\frac{i-1}{l_k-1} - \frac{j-1}{l_b-1}\right|^2 + \left|\frac{l_k-i}{l_k-1} - \frac{l_b-j}{l_b-1}\right|^2\right] \\ &= \mathbb{E}_{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}}\left[\left|\frac{i-1}{l_k-1} - \frac{j-1}{l_b-1}\right|^2 + \left|-\left(\frac{i-1}{l_k-1} - \frac{j-1}{l_b-1}\right)\right|^2\right] \\ &= 2\,\mathbb{E}_{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}}\left[\left|\frac{i-1}{l_k-1} - \frac{j-1}{l_b-1}\right|^2\right].\end{aligned} \tag{13}$$

Note that we apply a scaling factor of 2 here to account for both the anchor points $y_{k,1}^{(c)}$ and $y_{k,l_k}^{(c)}$ in the linearization of GW_2. Substituting GW_2 in Eq. (12) by $\mathrm{GW}_2^{\mathrm{linear}}$ in Eq. (13) yields

$$\mathrm{FGW}_{2,\alpha}^{\mathrm{linear}} = \min_{\mathbf{T}\in\Pi(\mu_k,\mu_b)} \mathbb{E}_{(y_{k,i}^{(c)},y_{b,j}^{(c)})\sim\mathbf{T}}\left[(1-\alpha)g^2\left(y_{k,i}^{(c)},y_{b,j}^{(c)}\right) + 2\alpha\left|\frac{i-1}{l_k-1} - \frac{j-1}{l_b-1}\right|^2\right] \tag{14}$$

which give a Wasserstein distance formulation that is linear w.r.t $\mathbf{T}$, thus proving the Lemma. □

A.2 Proof of Theorem 3.2

Theorem (Approximation Error of Anchor-Guided Linearization)**.** *Let $\mathbf{T}^*_{\mathrm{FGW}} \in \Pi(\mu_k, \mu_b)$ be an optimal transport plan that minimizes the original FGW problem with uniform marginal distributions μ_k and μ_b. The approximation error of the anchor-guided linearization in Lemma 3.1 is bounded as follows*

$$0 \le \mathrm{FGW}^{\textit{linear}}_{2,\alpha}(\mathcal{G}_k^{(c)}, \mathcal{G}_b^{(c)}) - \mathrm{FGW}_{2,\alpha}(\mathcal{G}_k^{(c)}, \mathcal{G}_b^{(c)}) \le 4\alpha\, \mathcal{C}(\mathbf{T}^*_{\mathrm{FGW}}), \tag{15}$$

where

$$\mathcal{C}(\mathbf{T}) = \mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[\frac{|I-I'|}{L_k-1}\frac{|J-J'|}{L_b-1}\mathbb{1}\{(I-I')(J-J')<0\}\right], \qquad (I,J),(I',J') \overset{\text{i.i.d.}}{\sim} \mathbf{T}. \tag{16}$$

Proof. Let $F(\mathbf{T})$ and $F_{\mathrm{lin}}(\mathbf{T})$ denote the original FGW formulation and its anchor-guided linearization evaluated at a common transport plan $\mathbf{T} \in \Pi(\mu_k, \mu_b)$, i.e.,

$$\begin{aligned} F(\mathbf{T}) &= (1-\alpha)\mathrm{W}_2(\mathbf{T}, \mathbf{M}_{k,b}) + \alpha\, \mathrm{GW}_2(\mathbf{T}, \mathbf{C}_k, \mathbf{C}_b), \\ F_{\mathrm{lin}}(\mathbf{T}) &= (1-\alpha)\mathrm{W}_2(\mathbf{T}, \mathbf{M}_{k,b}) + 2\alpha \sum\nolimits_{i=1,j=1}^{L_k,L_b} \left(\frac{i-1}{L_k-1} - \frac{j-1}{L_b-1}\right)^2 \mathbf{T}(i,j). \end{aligned} \tag{17}$$

Based on $(I,J),(I',J')$ sampled i.i.d. from $\mathbf{T}$, define the normalized token locations P, Q, P', and Q' as follows

$$P = \frac{I-1}{L_k-1}, \quad Q = \frac{J-1}{L_b-1}, \quad P' = \frac{I'-1}{L_k-1}, \quad Q' = \frac{J'-1}{L_b-1}.$$

In this way, we can rewrite $\mathrm{GW}_2(\mathbf{T}, \mathbf{C}_k, \mathbf{C}_b)$ and its linearization in Eq. (17) as follows

$$\begin{aligned} \mathrm{GW}_2(\mathbf{T}, \mathbf{C}_k, \mathbf{C}_b) &= \sum_{i,i'=1}^{L_k}\sum_{j,j'=1}^{L_b}\left(\frac{|i-i'|}{L_k-1} - \frac{|j-j'|}{L_b-1}\right)^2 \mathbf{T}(i,j)\mathbf{T}(i',j') \\ &= \mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[\left(|P-P'| - |Q-Q'|\right)^2\right] \end{aligned} \tag{18}$$

$$\sum_{i=1,j=1}^{L_k,L_b}\left(\frac{i-1}{L_k-1} - \frac{j-1}{L_b-1}\right)^2 \mathbf{T}(i,j) = \mathbb{E}_{\mathbf{T}}\left[(P-Q)^2\right]. \tag{19}$$

For any $a, b \in \mathbb{R}$, we have

$$(|a|-|b|)^2 = (a-b)^2 - 4|ab|\mathbb{1}\{ab<0\}. \tag{20}$$

Applying Eq. (20) with $a = P-P'$ and $b = Q-Q'$ gives

$$\begin{aligned} &\mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[\left(|P-P'| - |Q-Q'|\right)^2\right] \\ &= \mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[\left((P-Q) - (P'-Q')\right)^2\right] - 4\mathcal{C}(\mathbf{T}). \end{aligned} \tag{21}$$

where

$$\mathcal{C}(\mathbf{T}) = \mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[|P-P'||Q-Q'|\mathbb{1}\{(P-P')(Q-Q')<0\}\right]. \tag{22}$$

Since both (P,Q) and (P',Q') are i.i.d., we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[\left((P-Q) - (P'-Q')\right)^2\right] \\ &\quad= \mathbb{E}_{\mathbf{T}}\left[(P-Q)^2\right] + \mathbb{E}_{\mathbf{T}}\left[(P'-Q')^2\right] - 2\left(\mathbb{E}_{\mathbf{T}}[P-Q]\right)^2 \\ &\quad= 2\mathbb{E}_{\mathbf{T}}\left[(P-Q)^2\right] - 2\left(\mathbb{E}_{\mathbf{T}}[P] - \mathbb{E}_{\mathbf{T}}[Q]\right)^2. \end{aligned} \tag{23}$$

Since $\mathbf{T} \in \Pi(\mu_k, \mu_b)$ where μ_k and μ_b are two uniform distributions, we have

$$\mathbb{E}[P] = \frac{1}{L_k}\sum_{i=1}^{L_k}\frac{i-1}{L_k-1} = \frac{1}{2}, \qquad \mathbb{E}[Q] = \frac{1}{L_b}\sum_{j=1}^{L_b}\frac{j-1}{L_b-1} = \frac{1}{2}. \tag{24}$$

Therefore, we have

$$\mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[((P-Q)-(P'-Q'))^2\right] = 2\mathbb{E}_{\mathbf{T}}\left[(P-Q)^2\right]. \tag{25}$$

Substituting Eq. (25) into Eq. (21) and (18) yields

$$\begin{aligned}\mathrm{GW}_2(\mathbf{T},\mathbf{C}_k,\mathbf{C}_b) &= \mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[(|P-P'|-|Q-Q'|)^2\right] \\ &= 2\mathbb{E}_{\mathbf{T}}\left[(P-Q)^2\right] - 4\mathcal{C}(\mathbf{T})\end{aligned} \tag{26}$$

Comparing Eq. (26) with Eq. (17) and (19) yields that, for every $\mathbf{T}\in\Pi(\mu_k,\mu_b)$, we have

$$F_{\mathrm{lin}}(\mathbf{T}) - F(\mathbf{T}) = 4\alpha\,\mathcal{C}(\mathbf{T}) \geq 0. \tag{27}$$

Let

$$\mathbf{T}^*_{\mathrm{FGW}} = \underset{\mathbf{T}\in\Pi(\mu_k,\mu_b)}{\arg\min}\ F(\mathbf{T}), \qquad \mathbf{T}^*_{\mathrm{lin}} = \underset{\mathbf{T}\in\Pi(\mu_k,\mu_b)}{\arg\min}\ F_{\mathrm{lin}}(\mathbf{T})$$

be optimal transport plans that minimize $F(\mathbf{T})$ and $F_{\mathrm{lin}}(\mathbf{T})$, we have

$$\mathrm{FGW}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) = F(\mathbf{T}^*_{\mathrm{FGW}}), \qquad F(\mathbf{T}^*_{\mathrm{FGW}}) \leq F(\mathbf{T}^*_{\mathrm{lin}}) \tag{28}$$

$$\mathrm{FGW}^{\mathrm{linear}}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) = F_{\mathrm{lin}}(\mathbf{T}^*_{\mathrm{lin}}), \qquad F_{\mathrm{lin}}(\mathbf{T}^*_{\mathrm{lin}}) \leq F_{\mathrm{lin}}(\mathbf{T}^*_{\mathrm{FGW}}) \tag{29}$$

For the lower bound, from Eq. (27) to (29), we have

$$\mathrm{FGW}^{\mathrm{linear}}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) = F_{\mathrm{lin}}(\mathbf{T}^*_{\mathrm{lin}}) \geq F(\mathbf{T}^*_{\mathrm{lin}}) \geq F(\mathbf{T}^*_{\mathrm{FGW}}) = \mathrm{FGW}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) \tag{30}$$

Therefore, we have

$$\mathrm{FGW}^{\mathrm{linear}}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) - \mathrm{FGW}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) \geq 0. \tag{31}$$

For the upper bound, from Eq. (27) to (29), we have

$$\begin{aligned}\mathrm{FGW}^{\mathrm{linear}}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) &= F_{\mathrm{lin}}(\mathbf{T}^*_{\mathrm{lin}}) \\ &\leq F_{\mathrm{lin}}(\mathbf{T}^*_{\mathrm{FGW}}) \\ &= F(\mathbf{T}^*_{\mathrm{FGW}}) + 4\alpha\mathcal{C}(\mathbf{T}^*_{\mathrm{FGW}}), \\ &= \mathrm{FGW}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) + 4\alpha\mathcal{C}(\mathbf{T}^*_{\mathrm{FGW}}).\end{aligned} \tag{32}$$

Therefore, we have

$$\mathrm{FGW}^{\mathrm{linear}}_{2,\alpha} - \mathrm{FGW}_{2,\alpha} \leq 4\alpha\mathcal{C}(\mathbf{T}^*_{\mathrm{FGW}}). \tag{33}$$

Combining Eqs. (31) and (33) yields

$$0 \leq= \mathrm{FGW}^{\mathrm{linear}}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) - \mathrm{FGW}_{2,\alpha}(\mathcal{G}_k^{(c)},\mathcal{G}_b^{(c)}) \leq 4\alpha\mathcal{C}(\mathbf{T}^*_{\mathrm{FGW}}). \tag{34}$$

Finally, since

$$P-P' = \frac{I-I'}{L_k-1}, \qquad Q-Q' = \frac{J-J'}{L_b-1},$$

replacing $P-P'$ and $Q-Q'$ in Eq. (22) yields

$$\mathcal{C}(\mathbf{T}) = \mathbb{E}_{\mathbf{T}\otimes\mathbf{T}}\left[\frac{|I-I'|}{L_k-1}\frac{|J-J'|}{L_b-1}\mathbb{1}\{(I-I')(J-J')<0\}\right].$$

In this way, we have proven Theorem 3.2. □

A.3 Proof of Theorem 3.3

Theorem (Complexity of HARA)**.** *The time complexity of* HARA *is* $\mathcal{O}(BKLL_b(d+S))$*, and the space complexity of* HARA *is* $\mathcal{O}(KLL_b + KLd + L_bd)$*, where* K *is the number of sampled rollouts of length* L*,* L_b *is the barycenter size,* d *is the hidden state dimension.* S *and* B *are the number of Sinkhorn and barycenter optimization iterations for solving Eq.* (6)*, respectively.* [4].

Proof. **For time complexity**, the computation of HARA consists mainly of two alternating optimization steps (Zeng et al., 2024): for each iteration, (1) solving Eq. (6) for all K Wasserstein OT problems, and (2) updating $\mathbf{X}_b^{(c)}$ based on the solved Wasserstein distance at the current iteration. For solving the Wasserstein distance for one rollout in Eq. (6), constructing the cost matrix requires $\mathcal{O}(LL_bd)$ and $\mathcal{O}(LL_b)$ time for computing the pairwise distances of the d-dimensional hidden states and normalized locations, respectively. By entropy regularization, each Wasserstein OT problem can be solved in $\mathcal{O}(SLL_b)$ time via the Sinkhorn algorithm (Gabriel & Marco, 2019). Therefore, solving Eq. (6) for all K rollouts costs

$$K \cdot \mathcal{O}(LL_bd + LL_b + SLL_b) = \mathcal{O}(KLL_b(d+S))$$

per iteration. Updating $\mathbf{X}_b^{(c)}$ requires $\mathcal{O}(KLL_bd)$ time. Therefore, over B barycenter optimization iterations, the total time complexity is

$$B \cdot \mathcal{O}(KLL_b(d+S) + KLL_bd) = \mathcal{O}(BKLL_b(d+S))\,.$$

For space complexity, storing the transport plan matrices $\mathbf{T}_k^{(c)} \in \mathbb{R}^{L\times L_b}$ for K Wasserstein OT problems requires $\mathcal{O}(KLL_b)$ space. Storing the hidden states of K rollouts and the barycenter requires $\mathcal{O}(KLd)$ and $\mathcal{O}(L_bd)$ space, respectively. Therefore, the total space complexity is

$$\mathcal{O}(KLL_b + KLd + L_bd)\,.$$

□

B Detailed Experimental Setup

B.1 Dataset Setup

Training Dataset. We train all models on the MATH dataset (Hendrycks et al., 2021) which contains 12,500 problems from mathematics competitions. We use the publicly available version at `https://huggingface.co/datasets/DigitalLearningGmbH/MATH-lighteval`. Following (Lightman et al., 2024), we exclude problems inside MATH500 from the MATH dataset, and use the remaining 12,000 problems for training.

Evaluation Datasets. We evaluate the trained models on five standard mathematical reasoning benchmarks, including MATH500, AMC23, AIME24, AIME25 (Lightman et al., 2024), and Minerva Math (Lewkowycz et al., 2022), all of which are publicly available at `https://huggingface.co/math-ai/datasets`.

B.2 Training Setup

Across all group-based RLVR methods, we remove the KL penality by setting $\beta = 0$ in the loss function (Liu et al., 2025; Yu et al., 2026), and set the number of sampled rollouts to 16 for each prompt during training for advantage estimation. For rollout generation, we set a temperature to 1.0 with $top_p = 1.0$ and $top_k = -1$. For evaluation, we set a temperature to 0.6 with $top_p = 0.7$ and the same top_k values. All evaluation metrics reported are averaged over 5 different runs. We set a training batch size of 128 and a mini-batch size of 16 with a maximum response length of 8196 tokens. For the optimizer, we use AdamW with a learning rate of $1e-6$, optimizer parameters of $\beta = (0.9, 0.999)$, and a weight decay of 0.01. For GRPO (Shao et al., 2024) and Dr.GRPO (Liu et al., 2025), we set the clipping parameter $\epsilon = 0.2$; for DAPO (Yu et al., 2026), we set the clipping parameter $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$. All experiments are conducted on 8 NVIDIA A100 GPUs.

[4] Without loss of generality, we assume $\mathcal{O}(K^{(+)}) \approx \mathcal{O}(K^{(-)}) \approx \mathcal{O}(K)$ and $\mathcal{O}(L_k^{(c)}) \approx \mathcal{O}(L)$ for all $k \in [1, K]$ and $c \in \{+, -\}$.

B.3 HYPERPARAMETER SETUP

For hyperparameters of HARA we set $\alpha = 0.25$ in Eq. (3), and the barycenter size $l_b = 32$. For the entropy regularization, we set the its coefficient to 0.05. Such hyperparameter setup is consistent across all experiments.

B.4 SETUP FOR INTERPRETABILITY STUDY IN FIGURE 2

In the interpretability study in Figure 2a and 2b, we set the number of sampled rollouts within each outcome-grounded group to $K = 16$ for computing the barycenters. The barycenter sizes of all OT variants are set to $l_b = 32$, and the entropy regularization coefficient is set to 0.05. For the FGW formulation, the hyperparameter α is set to 0.25. All experiments in Figure 2 are conducted on the Qwen3-8B-Base (Yang et al., 2025). The prompt for GPT-5-Sol to derive the reasoning stages of each rollout and classify tokens to these stages is included below.

Structured Reasoning-Stage Annotation Prompt

You are given one problem and multiple reasoning rollouts.
Your task is to infer a **shared superset of abstract reasoning stages** that collectively explains the reasoning performed across all provided rollouts, and then assign every token in every rollout to one of these stages.
The stages should describe **reasoning functions or roles**, rather than surface wording. Examples include understanding or reformulating the problem, establishing an intermediate result, applying a theorem or relation, case analysis, algebraic simplification, verification, and deriving the final answer.

Shared Stage Superset

Infer a stage vocabulary

$$\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$$

that is sufficiently expressive to accommodate the reasoning content appearing anywhere in the rollout group.
The stage vocabulary is a **superset**: a stage does not need to appear in every rollout.
For example, if some rollouts solve the problem directly while others introduce an auxiliary lemma, the shared vocabulary should contain an `Auxiliary Lemma` stage even though the direct rollouts skip it.
Different rollouts may therefore:

- skip one or more stages;
- repeat a stage;
- use only a subset of the shared stages;
- realize the same stage using substantially different wording;
- execute shared stages in different orders when their reasoning genuinely differs.

Do **not** require every rollout to instantiate every stage.
At the same time, avoid unnecessarily creating rollout-specific stages. If two spans perform the same underlying reasoning function despite different wording or mathematical expressions, assign them to the same abstract stage.

Complete Token Coverage

Every token in every rollout must be assigned to one of the inferred reasoning stages.
Do not use `OTHER`, `UNKNOWN`, or any unassigned category.
If some tokens do not fit the current stage vocabulary, refine or expand the shared stage vocabulary so that they can be assigned to an appropriate reasoning stage.
This includes transitional language, equations, mathematical symbols, intermediate calculations, verification statements, and final-answer text. Assign such tokens to the reasoning stage whose function they support.
Prefer a compact and meaningful stage vocabulary over creating stages for individual words or superficial formatting differences.

Annotation Procedure

1. Read the problem and **all rollouts jointly** before defining the stages.
2. Identify the union of distinct reasoning functions required to explain the rollouts.
3. Consolidate semantically equivalent reasoning functions into shared stages even when their surface realizations differ.
4. Create additional stages when necessary to accommodate reasoning strategies that occur only in a subset of rollouts.

5. For each rollout, segment the complete token sequence into contiguous spans.
6. Assign every span to exactly one stage from the shared stage vocabulary.
7. Preserve genuine differences in reasoning order. Do not relabel spans merely to make all rollouts follow the same stage sequence.

Output Requirements

For each stage, provide:

- `stage_id`
- `stage_name`
- `definition`

For each rollout, provide contiguous token spans with:

- `start_token`
- `end_token`
- `stage_id`

Use the supplied token indices exactly. Do not invent, modify, or renumber tokens.
A stage may occur multiple times in a rollout, in which case multiple spans may have the same `stage_id`.

Output Format

Return **valid JSON only** using the following schema:

```
{
    "stages": [
        {
            "stage_id": "C1",
            "stage_name": "...",
            "definition": "..."
        },
        {
            "stage_id": "C2",
            "stage_name": "...",
            "definition": "..."
        }
    ],
    "rollouts": [
        {
            "rollout_id": "R1",
            "spans": [
                {
                    "start_token": 0,
                    "end_token": 12,
                    "stage_id": "C1"
                },
                {
                    "start_token": 13,
                    "end_token": 27,
                    "stage_id": "C3"
                }
            ]
        }
    ]
}
```

Final Validation

Before returning the JSON, verify all of the following:

- Every rollout token is assigned exactly once.
- The first span begins at the first token.
- The final span ends at the final token.
- Consecutive spans are contiguous with no gaps.
- Spans never overlap.
- Every assigned `stage_id` exists in `stages`.
- No `OTHER`, `UNKNOWN`, or unlabeled tokens are present.
- Stages form a **superset across the rollout group**; individual rollouts are allowed to skip stages.
- Stages capture reasoning functions rather than superficial wording, token type, or formatting.
- The stage vocabulary is no larger than necessary to explain meaningful reasoning differences across the rollouts.

Return JSON only.

Table 3: Additional benchmark results of credit assignment methods on Dr.GRPO and DAPO with Qwen3-8B-Base. The 1st/2nd best results are highlighted in **bold**/underline, respectively. Our methods are highlighted in red box.

Method	MATH500	Minerva Math	AMC23	AIME24	AIME25	Average
			Qwen3-8B-Base			
Dr.GRPO	80.1 ±3.1	33.8 ±2.5	87.0 ±1.7	53.3 ±0.9	42.0 ±0.6	59.2
+ GRPO-λ	81.5 ±1.7	34.1 ±0.3	87.0 ±0.9	50.7 ±0.6	39.3 ±0.9	58.5
+ Entropy Adv.	82.8 ±2.8	35.4 ±1.9	**88.5** ±0.7	52.7 ±0.6	44.7 ±1.3	60.8
+ SHEAR	82.6 ±1.5	35.9 ±2.1	86.0 ±1.3	54.7 ±1.3	**45.3** ±1.1	60.9
+ HARA-seq	83.2 ±1.4	35.2 ±1.0	87.5 ±1.1	55.3 ±1.1	44.7 ±0.6	61.2
+ HARA-tok	**84.3** ±1.8	**36.1** ±1.2	**88.5** ±0.4	**56.7** ±0.6	44.7 ±0.9	**62.1**
			Qwen3-8B-Base			
DAPO	82.1 ±2.8	35.8 ±2.7	89.0 ±0.9	52.0 ±1.1	42.0 ±0.9	60.2
+ GRPO-λ	81.8 ±1.9	35.2 ±1.7	88.0 ±0.9	51.3 ±0.6	40.7 ±0.6	59.4
+ Entropy Adv.	83.9 ±2.0	36.1 ±1.9	90.5 ±1.1	52.7 ±0.9	42.7 ±1.1	61.2
+ SHEAR	82.7 ±1.6	**37.2** ±1.3	90.0 ±1.4	55.3 ±0.6	44.0 ±0.6	61.9
+ HARA-seq	**85.2** ±1.7	36.5 ±2.1	89.5 ±0.4	**56.7** ±0.6	44.7 ±0.6	62.5
+ HARA-tok	85.1 ±1.0	37.1 ±0.9	**91.5** ±1.3	**56.7** ±0.9	**45.3** ±0.6	**63.1**

C Additional Experiment Results

C.1 Additional Benchmark Results

To further verify the generalizability of HARA on different group-based RLVR methods, we conduct additional benchmark experiments on Dr.GRPO (Liu et al., 2025) and DAPO (Yu et al., 2026) using Qwen3-8B-Base. The results are shown in Table 3, which demonstrate that **1)** HARA consistently improves the performance of Dr.GRPO and DAPO across all datasets, and **2)** HARA predominantly outperforms existing credit assignment methods on most datasets and LLMs. Both observations demonstrate the effectiveness and generalizability of HARA in improving group-based RLVR methods by identifying and emphasizing novel reasoning elements during RL.

C.2 Additional Studies

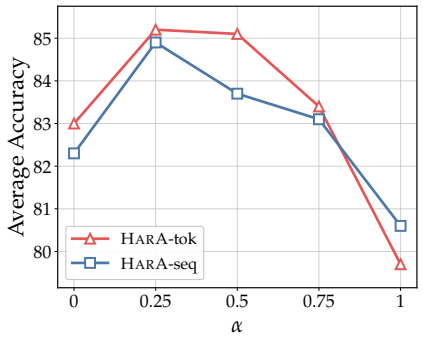


Figure 7: Parameter study on α by Qwen3-8B-Base on MATH500.

Table 4: Ablation study of anchor-guided linearization by Qwen3-8B-Base on MATH500.

Method	MATH500	
	Avg. Accuracy	Time / Step
HARA(w/o AGL)-seq	84.4	154
HARA(w AGL)-seq	84.7	41.0
Δ	−0.3	3.75×
HARA(w/o AGL)-tok	85.3	175
HARA(w AGL)-tok	85.2	43.6
Δ	+0.1	4.01×

Parameter Study on α. We conduct a parameter study on the parameter α in Eq. (1) which controls the trade-off between the Wasserstein distance and Gromov-Wasserstein distance in the FGW formulation. The results on MATH500 with Qwen3-8B-Base by GRPO are shown in Figure 7, where we can see that **the performance of HARA remain relatively stable when α is in the range of $[0.25, 0.75]$**, with less than 2% performance variation. If $\alpha = 0$ or 1, the FGW formulation reduces to the Wasserstein distance or Gromov-Wasserstein distance, respectively, which fails to

capture either the semantic or structural information of a rollout, leading to more severe performance degradation. We set $\alpha = 0.25$ as the default value in HARA.

Ablation on Anchor-Guided Linearization. To further understand how the anchor-guided linearization in Lemma 3.1 affects the empricial performance and efficiency of HARA, we conduct an ablation study on MATH500 with Qwen3-8B-Base by GRPO, comparing the performance of HARA with w/ and w/o the anchor-guided linearization (AGL), as shown in Table 4. We can observe that **1) the anchor-guided linearization hardly affect the performance of** HARA, with less than 1% performance variation. Such observation is consistent with our theoretical analysis in Lemma 3.1, where we show that the approximation error of the linearization is tightly bounded. Meanwhile, **2) the anchor-guided linearization significantly improves the efficiency of** HARA thanks to strict convexity of the linearized problem Gabriel & Marco (2019), with up to $4\times$ speed-up in training time per gradient update step. This futher justify using the anchor-guided linearization in HARA to speed-up the OT computation in HARA.