GRADIENT-SPACE DIVERSITY AS CREDIT ASSIGNMENT FOR POST-TRAINING LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning has become a standard approach for post-training large language models, but improvements in task performance often come with reduced output diversity and increasingly narrow solution patterns. Most existing diversity-aware training methods quantify variation in output, embedding, or hidden-state space and optimize the resulting metric as an auxiliary reward. In contrast, this work leverages training gradients as the primary signal to measure and promote diversity. First, under a local first-order approximation, we show that GRPO propagates group-relative advantages through the rollout-gradient Gram matrix, exposing gradient redundancy among verified solutions as a hidden dimension of credit assignment. Motivated by this observation, we introduce **GraDiC**, a novel training method that directly operates on this gradient space. GraDiC measures their coverage using the order-2 Vendi Score and differentiates this objective with respect to each response's update weight, yielding a closed-form influence score that quantifies the response's marginal contribution to gradient-space diversity. The resulting update mechanism promotes underrepresented gradient directions and mitigates the tendency of policy updates to concentrate along dominant directions. Experiments on challenging mathematical reasoning benchmarks show that GraDiC substantially improves pass@k for $k \in [16, 128]$ compared with standard baselines.

1 INTRODUCTION

Post-training with reinforcement learning has become a central paradigm for improving large language models, adapting their behavior to user instructions and task requirements while extending their capabilities beyond supervised fine-tuning (Ouyang et al., 2022; Rafailov et al., 2023; Bai et al., 2022). However, a major challenge in reinforcement training is exploration (Ladosz et al., 2022; Gupta et al., 2018). While RL improves response quality, standard algorithms often cause the model to collapse into a limited set of canonical responses (Yue et al., 2025). This sharp reduction in output diversity compared to the base model has several negative consequences, particularly when a prompt admits multiple valid perspectives or solution strategies. It restricts personalization, weakens test-time search capabilities, and reduces robustness under distribution shift (Anderson et al., 2024). Furthermore, during training on reasoning tasks, a narrow response distribution restricts the paths explored, making it harder for models to discover novel and potentially more effective strategies. Ultimately, these challenges motivate the need for post-training objectives that can consistently improve response quality while retaining a broad range of useful outputs.

A growing body of work has therefore sought to preserve exploration and diversity during LLM post-training. One line of work encourages additional stochasticity in the policy, for example through token-level entropy or diversity objectives on successful trajectories (Yao et al., 2025). Other approaches construct exploration rewards from output-level notions of novelty, such as structural and lexical differences across sampled solutions (Hu et al., 2026; Zhou et al., 2025). More recently, diversity has been measured in semantic representation spaces: DQO uses determinantal point processes over response embeddings (Chen et al., 2026), while SEED-GRPO uses semantic entropy over sampled responses to modulate group-relative updates (Chen et al., 2025). Outcome-based exploration provides a complementary perspective by explicitly rewarding underrepresented suc-

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.

cessful answers (Song et al., 2025). Despite these different formulations, most existing diversity signals are defined over lexical or semantic variation in generated responses, which provides only a coarse view of how individual responses differ in their effect on model updates.

A more direct measure of diversity must therefore emerge from the optimization dynamics. Since output-level signals cannot determine whether seemingly diverse solutions induce redundant update directions—or, conversely, whether semantically indistinguishable responses yield distinct gradients—this naturally motivates using parameter gradients to evaluate diversity in the update space.

A notable recent step in this direction is G^{2}RL (Liang et al., 2025), which constructs an exploration reward from output-layer gradient features. However, this approach has two key limitations. First, it relies on a heuristic diversity metric over pairwise similarities of logit-layer gradients. Second, because these output-layer gradients maintain a closed-form dependence on hidden states and output probabilities, the resulting representation remains closely tied to standard semantic features. Consequently, it remains an open question how actual gradient-space diversity can be defined and translated into response-level credit in a principled manner.

To address this gap, we introduce GRADIC, which measures diversity directly over the parameter gradients of a full transformer layer. Motivated by our empirical finding that middle layers undergo the most pronounced parameter changes during RL training, this formulation effectively captures the core optimization dynamics. We develop an efficient implementation within the VeRL framework, adding merely an 11.6% runtime overhead compared to standard GRPO. Beyond that, GRADIC rests on a principled theoretical foundation. Specifically, we show under a local first-order approximation that GRPO propagates group-relative advantages through the rollout-gradient Gram matrix, revealing gradient redundancy among verified solutions as an implicit dimension of credit assignment. GRADIC quantifies coverage of these gradient directions using the order-2 Vendi Score and differentiates this set-level objective with respect to each response's update weight, yielding a closed-form marginal contribution to gradient-space diversity. The resulting update rule favors underrepresented gradient directions over already dominant ones, counteracting the concentration of policy updates around redundant solutions.

Together, our main contributions are:

- We provide a local first-order analysis of GRPO showing that sampled-response interactions are governed by the rollout-gradient Gram matrix, exposing gradient redundancy as an additional dimension of credit assignment.
- We introduce **GraDiC**, which defines diversity over full transformer layer gradients and derives a closed-form per-response credit signal by differentiating an order-2 Vendi objective with respect to each response's update weight.
- We characterize the proposed credit mechanism through its connections to gradient-space coverage, local changes in response probabilities, and transfer under an ℓ_2 uncertainty model. The resulting update favors underrepresented gradient directions over dominant ones that lead to collapse.
- We develop an efficient implementation that incurs only $1.116\times$ the runtime of GRPO. Experiments on challenging mathematical reasoning benchmarks show that GRADIC substantially improves the solution coverage measured by pass@k.

2 PRELIMINARIES

2.1 GRPO WITH VERIFIABLE REWARDS

The standard GRPO objective (ignoring KL-divergence term) is:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(\rho_{i,t}(\theta)A_i, \mathrm{clip}\left(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)A_i\right), \tag{1}$$

where notation follows the standard GRPO setup; definitions are summarized in Appendix A.1. Ignoring the KL term, at $\theta = \theta_{\mathrm{old}}$, $\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t}|q,o_{i,<t})} = 1$ the update gradient is:

$$-\eta \, \nabla_\theta \mathcal{L}_{\mathrm{GRPO}}(\theta)|_{\theta=\theta_{\mathrm{old}}} = \frac{\eta}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} A_i \sum_{t=1}^{T_i} \nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t})|_{\theta=\theta_{\mathrm{old}}}, \tag{2}$$

Per-response Gradient. The contribution of response o_i to the update is the length-normalized summation of token-level gradients:

$$\eta \frac{1}{|o_i|} A_i \, \nabla_\theta \log \pi_\theta(o_i \mid q)|_{\theta=\theta_{\mathrm{old}}} = \eta \frac{1}{|o_i|} A_i \sum_{t=1}^{|o_i|} \nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t})|_{\theta=\theta_{\mathrm{old}}}, \tag{3}$$

where we omit normalization factors shared across responses. When comparing responses with the same advantage, as η and A_i are the same, their relative effect on the update is determined by the *geometry of their response gradients*, $\nabla_\theta \log \pi_\theta(o_i \mid q)|_{\theta=\theta_{\mathrm{old}}}$.

2.2 The Vendi Score Family

The Vendi Score is a similarity-based measure of diversity that quantifies the effective number of distinct elements in a collection (Friedman & Dieng, 2023). Unlike diversity measures that require a predefined partition into classes or a reference distribution, the Vendi Score operates directly on pairwise similarities, allowing the relevant notion of diversity to be specified through a user-defined similarity function.

Specifically, consider a collection of n elements $\mathcal{X} = \{x_1, \ldots, x_n\}$ and a positive semidefinite similarity function $k(\cdot, \cdot)$ satisfying $k(x_i, x_i) = 1$. Let $S_{ij} = k(x_i, x_j)$ and $\widetilde{S} = \frac{1}{n} S$, with $\lambda_1, \ldots, \lambda_n$ denoting the eigenvalues of $\widetilde{S}$. Since $\mathrm{tr}(\widetilde{S}) = 1$, $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. The original Vendi Score is the exponential of the Shannon entropy of this eigenspectrum:

$$\mathrm{VS}_1(\mathcal{X}; k) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i \right). \tag{4}$$

Thus, $\mathrm{VS}_1(\mathcal{X}; k) \in [1, n]$, where $\mathrm{VS}_1 = n$ if $S = I_n$ (elements are mutually orthogonal) and $\mathrm{VS}_1 = 1$ if $S = \mathbf{1}\mathbf{1}^\top$ (elements are identical).

Vendi Score Family. Broadly, the original Vendi Score corresponds to a particular member of a broader family of similarity-based diversity measures obtained by replacing Shannon entropy with Rényi entropy (Pasarkar & Dieng, 2024). For order $\alpha \geq 0$ and $\alpha \neq 1$, the Vendi Score is:

$$\mathrm{VS}_\alpha(\mathcal{X}; k) = \left(\sum_{i : \lambda_i > 0} \lambda_i^\alpha \right)^{\frac{1}{1-\alpha}}, \tag{5}$$

with VS_1 defined by the continuous limit in Eq. 4. Several orders admit useful interpretations:

$$\mathrm{VS}_0 = \mathrm{rank}(\widetilde{S}), \quad \mathrm{VS}_1 = \exp\left(-\sum_i \lambda_i \log \lambda_i \right), \quad \mathrm{VS}_2 = \frac{1}{\mathrm{tr}(\widetilde{S}^2)}, \quad \mathrm{VS}_\infty = \frac{1}{\lambda_{\max}(\widetilde{S})}. \tag{6}$$

Here, the order α controls the metric's sensitivity to the input: lower orders emphasize the presence of rare, low-prevalence spectral modes (but also risk amplifying noisy inputs), whereas higher orders emphasize common elements and is more resistant to minor intra-class variance or noise.

Of particular interest is the order-2 score, which admits a direct collision interpretation. For uniformly weighted samples,

$$\mathrm{tr}(\widetilde{S}^2) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j)^2, \tag{7}$$

meaning VS_2 is the inverse average squared similarity between two independently drawn elements:

$$\mathrm{VS}_2 = \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j)^2 \right)^{-1}. \tag{8}$$

A high VS_α indicates that similarity mass is spread across many distinct directions rather than concentrated in a small number of redundant modes. The Vendi construction is agnostic to the representation used to define similarity. In particular, when each element is represented by a normalized gradient vector and similarity is given by gradient inner products, with $\alpha = 1$, the resulting metric is referred to as *gradient Vendi* (G-Vendi) (Jung et al., 2025). In this case, spectral diversity measures the effective number of distinct gradient directions, providing a notion of diversity directly in the geometry of model updates. We further derive VS_2's connection to the update probability during GRPO training, as well as other properties, in Section 4

3 METHOD

3.1 EXTRACTING GRADIENT FEATURES

From a group of G rollouts, let $\mathcal{P} = \{i : A_i > 0\}$ denote the set of correct responses. For each $i \in \mathcal{P}$, we extract the length-normalized response gradient:

$$u_i = \frac{1}{|o_i|} \nabla_\theta \log \pi_\theta(o_i \mid q)|_{\theta=\theta_{\text{old}}}, \qquad i \in \mathcal{P}, \tag{9}$$

and its normalized direction as

$$g_i = \frac{u_i}{\|u_i\|_2}, \qquad \|g_i\|_2 = 1. \tag{10}$$

In practice, evaluating the gradient for the entire model parameter is prohibitively expensive. To address this, we compute the gradient with respect to the parameters of a single transformer layer. Following Xia et al. (2024), we further project these high-dimensional gradients down to $d = 1024 \ll |\theta|$ to efficiently collect the final gradient representation. This is achieved via a Rademacher random projection $\Pi \in \mathbb{R}^{|\theta| \times d}$, where $\Pi_{ij} \sim \mathcal{U}(\{-1, 1\})$, qualifying as a Johnson-Lindenstrauss transform that preserves the inner product between projected gradients with high probability.

3.2 WEIGHTED ORDER-2 GRADIENT VENDI

Update Gradient Decomposition. The positive side of the GRPO update aggregates the correct response gradients:

$$\Delta\theta_+ \propto \sum_{i \in \mathcal{P}} A_i u_i. \tag{11}$$

To decouple magnitude from geometry of the gradients, we define the per-response update mass $w_i = A_i \|u_i\|_2$ and the total mass $W_+ = \sum_{i \in \mathcal{P}} w_i$. This induces a probability distribution $p_i = w_i / W_+$ over the unit directions g_i, allowing us to rewrite the update as:

$$\Delta\theta_+ \propto W_+ \sum_{i \in \mathcal{P}} p_i g_i. \tag{12}$$

This decomposition isolates the total update scale W_+ from its directional distribution p, two quantities that standard GRPO implicitly couples.

Differentiable Weighted Vendi. The Vendi Score (Section 2.2) evaluates diversity over a set of uniformly weighted samples. Our key insight is to instead parameterize the score with the update masses w_i, treating them as *differentiable variables*. Transitioning from a static set-level metric to a differentiable objective allows us to perform response-level credit assignment. To this end, we define the weighted gradient second-moment matrix as:

$$K(w) = \frac{1}{W_+} \sum_{i \in \mathcal{P}} w_i g_i g_i^\top = \sum_{i \in \mathcal{P}} p_i g_i g_i^\top. \tag{13}$$

For notation simplicity, we hence write K in place of $K(w)$ whenever the dependence on w is clear from context. Since $\|g_i\|_2 = 1$ and $\sum_{i \in \mathcal{P}} p_i = 1$, we have $\mathrm{tr}(K(w)) = \sum_{i \in \mathcal{P}} p_i \|g_i\|^2 = 1$. The order-2 gradient Vendi Score is therefore defined as:

$$\mathrm{GV}_2(w) = \frac{1}{\mathrm{tr}(K(w)^2)}. \tag{14}$$

Interpretation. Expanding the trace gives a clear pairwise interpretation:

$$\mathrm{tr}(K(w)^2) = \sum_{i,j\in\mathcal{P}} p_i p_j \langle g_i, g_j\rangle^2. \tag{15}$$

Thus, GV_2 is the inverse expected squared collision between two independently drawn verified gradient directions under the current positive mass distribution p. A concentrated update, in which most mass lies on a few mutually aligned directions, has a low GV_2; conversely, a more evenly covered set of gradient directions has a high GV_2. We further elaborate on the theoretical connection between GV_2 and the rollout-gradient Gram matrix in the following section.

3.3 GRADIC: GRADIENT-BASED DIVERSITY CREDIT ASSIGNMENT

Deriving Vendi's Credit Assignment. To translate this differentiable set-level objective into a per-response credit signal, first, we define our objective as the logarithm of the order-2 Vendi Score:

$$\Phi(w) = \log \mathrm{GV}_2(w) = -\log \mathrm{tr}(K(w)^2). \tag{16}$$

Because the normalization $W_+ = \sum_j w_j$ also depends on w_i, the derivative of the covariance matrix with respect to the weight of response i is:

$$\frac{\partial K}{\partial w_i} = \frac{1}{W_+}\left(g_i g_i^\top - K\right). \tag{17}$$

Using the identity $\partial \,\mathrm{tr}(K^2) = 2\,\mathrm{tr}(K\,\partial K)$, we obtain the following closed-form influence for each response.

Proposition 1 (Order-2 Vendi influence)**.** *For every verified positive response $i \in \mathcal{P}$,*

$$\frac{\partial}{\partial w_i} \log \mathrm{GV}_2(w) = \frac{2}{W_+}\left(1 - \frac{g_i^\top K g_i}{\mathrm{tr}(K^2)}\right). \tag{18}$$

Response Credit. This formalizes the scale-free response credit D_i:

$$D_i := W_+ \frac{\partial}{\partial w_i} \log \mathrm{GV}_2(w) = 2\left(1 - \frac{c_i}{C}\right), \tag{19}$$

where

$$c_i := g_i^\top K g_i = \sum_{j\in\mathcal{P}} p_j \langle g_i, g_j\rangle^2, \qquad C := \mathrm{tr}(K^2) = \sum_{i\in\mathcal{P}} p_i c_i. \tag{20}$$

Here, c_i measures the gradient redundancy of response i with respect to the current positive-response distribution, while C is its weighted average across responses. Consequently, if $c_i < C$, response i is less redundant than the average verified response and receives positive Vendi credit. If $c_i > C$, its direction is overrepresented and it receives negative Vendi credit. Importantly, this credit is automatically centered under the current positive mass ($\sum_{i\in\mathcal{P}} p_i D_i = 0$). Hence, D_i measures a response's contribution to coverage relative to the current verified population, rather than assigning an absolute notion of quality.

Here, we note that other orders also admit per-response influence functions. We also derive the formulation of differentiating GV_1 for credit assignment in Appendix D. Our choice of GV_2 in the main method is because it is a sweet spot to balance signal and noise, as GV_1 tends to be more sensitive to noise, and GV_2 offers nice theoretical properties described in Section 4.

Advantage Reweighting. To apply these influence scores during training, we convert them into a new probability distribution over the verified positive responses using exponential tilting, yielding the updated mass $\widetilde{w}_i$:

$$\widetilde{p}_i = \frac{p_i \exp(\beta \widehat{D}_i)}{\sum_{j\in\mathcal{P}} p_j \exp(\beta \widehat{D}_j)}, \qquad \widetilde{w}_i = W_+ \widetilde{p}_i, \tag{21}$$

where $\beta \geq 0$ controls the strength of coverage reweighting (note that $\beta = 0$ recovers the original GRPO setting), and $\widehat{D}_i$ denotes the stop-gradient version of D_i. More discussion on the sensitivity of β is shown in Appendix E

By defining a multiplicative coefficient $\lambda_i = \frac{\widetilde{p}_i}{p_i}$, we replace the original positive GRPO advantage with $\widetilde{A}_i = \lambda_i A_i$. The resulting positive policy update becomes:

$$\sum_{i \in \mathcal{P}} \widetilde{A}_i u_i = W_+ \sum_{i \in \mathcal{P}} \widetilde{p}_i g_i. \tag{22}$$

Ultimately, GRADIC strictly redistributes a fixed positive update mass among verified responses based on the coverage of gradients. Negative responses retain their original advantages.

4 THEORETICAL PROPERTIES OF GRADIC

4.1 GRADIENT SIMILARITY COUPLES GRPO UPDATES

In this section, we derive how the log-probabilities of a group of sampled rollouts change after a local GRPO update, demonstrating that this shift is governed by the geometry of their gradients. Let $u_i = \frac{1}{|o_i|} \nabla_\theta \log \pi_\theta(o_i \mid q)|_{\theta=\theta_{\text{old}}}$ be the normalized gradient of rollout o_i. Let U be the matrix formed by stacking these full-parameter gradients evaluated at $\theta = \theta_{\text{old}}$:

$$U = \begin{bmatrix} u_1^\top \\ \vdots \\ u_G^\top \end{bmatrix}. \tag{23}$$

With the advantage vector $A = (A_1, \ldots, A_G)^\top$, and absorbing the learning rate and group-size normalization into a scalar τ, the reward-only GRPO gradient step is given by $\Delta\theta = \tau U^\top A$.

To understand how this parameter update affects the policy output, a first-order Taylor expansion of the length-normalized log-probability for each response around θ_{old} yields:

$$\frac{1}{|o_i|} \log \pi_{\theta_{\text{old}}+\Delta\theta}(o_i \mid q) \approx \frac{1}{|o_i|} \log \pi_{\theta_{\text{old}}}(o_i \mid q) + u_i^\top \Delta\theta. \tag{24}$$

Therefore, the first-order log-probability change for response i is approximated by $u_i^\top \Delta\theta$. Together, stacking these scalar changes yields the log-probability change vector of the rollout group:

$$\Delta \log \pi_G := \begin{bmatrix} \Delta \log \pi(o_1 \mid q) \\ \vdots \\ \Delta \log \pi(o_G \mid q) \end{bmatrix} \approx U \Delta\theta = \tau U U^\top A. \tag{25}$$

Equation 25 shows that response-level credit in GRPO is inherently coupled through the rollout-gradient Gram matrix $UU^\top$. In particular, the advantage A_j assigned to response j affects the first-order log-probability change of response i in proportion to their gradient alignment $\langle u_i, u_j \rangle$. Thus, responses with similar gradients receive strongly coupled updates, making gradient redundancy directly relevant to how learning signals are propagated within a sampled group.

4.2 GEOMETRIC INTERPRETATION OF GRADIENT CONCENTRATION

The Gram-matrix formulation naturally motivates measuring how the positive update mass concentrates along overlapping geometric directions. We retain the unit directions g_i and the mass distribution p from Section 3, where $K = \sum_i p_i g_i g_i^\top$. Let $V = \text{span}\{g_i : i \in \mathcal{P}\}$ and $r = \dim V$. Expressing K in orthonormal coordinates on V, the matrix I_r/r represents an equal distribution of second-moment mass across all r available axes.

Proposition 2 (GV_2 and directional concentration)**.** *For unit gradient directions and a probability distribution p,*

$$\left\| K - \frac{I_r}{r} \right\|_F^2 = \text{tr}(K^2) - \frac{1}{r} = \frac{1}{\text{GV}_2(p)} - \frac{1}{r}. \tag{26}$$

Thus, maximizing GV_2 is geometrically equivalent to pushing the weighted gradient second moment toward a uniform mass distribution across its span, steering the optimization toward ***a broader, more balanced gradient distribution*** that mitigates directional collapses.

4.3 Local Effect of the GraDiC Reweighting Mechanism

Let $p_i(\beta) \propto p_i \exp(\beta D_i)$ denote the reweighted distribution, with $p(0) = p$.

Theorem 3 (Local ascent of gradient coverage)**.** *For a fixed set of unit gradient features and their Vendi credits,*

$$\frac{d}{d\beta} \log \mathrm{GV}_2(p(\beta))\Big|_{\beta=0} = \mathrm{Var}_{i\sim p}(D_i) \geq 0. \tag{27}$$

Equality holds iff the credits are identically equal on the support of p. Otherwise, every sufficiently small positive step strictly increases GV_2.

Because $\sum_i p_i D_i = 0$, the variance simplifies to $\sum_i p_i D_i^2$, meaning a sufficiently small positive step **strictly increases** the GV_2 objective whenever responses exhibit differing levels of gradient redundancy. If all responses receive the same credit, then $D_i = 0$ for all i, and GRADIC leaves their relative update weights unchanged from standard GRPO .

4.4 Gradient Coverage and Out-of-Distribution Transfer

Since learned directions generalize unevenly under distribution shift, concentrated updates are vulnerable to directions that transfer poorly. We examine how update mass distribution affects transfer to an out-of-distribution (OOD). Suppose the correct-response directions belong to r orthonormal modes $e_1, \ldots, e_r$, with mass $\rho_c = \sum_{i:g_i=e_c} p_i$ on mode c. For an unseen target loss L_T, let $h_T = -\nabla L_T(\theta_{\mathrm{old}})$. Defining $\xi_c = h_T^\top e_c$, the first-order target transfer is proportional to $W_+ \rho^\top \xi$. Assuming a mean transfer $\bar{\xi}$ and bounded variance σ^2 across these modes, let $\mathcal{U}(\bar{\xi}, \sigma)$ denote the uncertainty set of transfer coefficients satisfying:

$$\frac{1}{r}\sum_c \xi_c = \bar{\xi}, \qquad \frac{1}{r}\sum_c (\xi_c - \bar{\xi})^2 \leq \sigma^2. \tag{28}$$

Theorem 4 (OOD transfer under mode uncertainty)**.** *In the orthonormal-mode model,* $\mathrm{GV}_2(\rho) = \left(\sum_c \rho_c^2\right)^{-1}$*, and*

$$\inf_{\xi \in \mathcal{U}(\bar{\xi},\sigma)} \rho^\top \xi = \bar{\xi} - \sigma\sqrt{\frac{r}{\mathrm{GV}_2(\rho)} - 1}. \tag{29}$$

Equation 29 shows that concentrating update mass on a few modes exposes the policy to poor transfer along dominant directions. Increasing GV_2 mitigates this vulnerability by spreading the update across more modes, thereby raising the guaranteed lower bound for target transfer.

5 Experiments

5.1 Setup

Models, data, and baselines. We evaluate Qwen2.5-Math-1.5B and Qwen2.5-Math-7B (Yang et al., 2024) using the DAPO-Math-17k dataset (Yu et al., 2025) as the training dataset. During training, we sample a group of 16 responses for each problem. We compare our approach against GRPO and standard baselines that cover a wide range of diversity training methods: entropy bonuses (R1-zero-Div (Yao et al., 2025)), heuristic rule-based rewards (DIVER (Hu et al., 2026)), semantic embeddings (DQO (Chen et al., 2026)), and the closest work to ours (G^{2}RL (Liang et al., 2025)).

Evaluation. We evaluate on five challenging olympiad-level mathematical reasoning benchmarks: AIME24, AIME25, AIME26, BeyondAIME (ByteDance-Seed, 2025), and BRUMO25 (Balunović et al., 2025). For each problem, we generate 256 responses and compute pass@k for $k \in [1, 128]$ (we exclude pass@256, as estimating it from a 256-sample pool leads to substantially higher variance):

$$\widehat{\text{pass@}}\, k = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}, \qquad 1 \leq k \leq n, \tag{30}$$

where c is the number of correct responses among n samples and $\binom{n-c}{k} = 0$ when $n - c < k$. Prompt templates, generation settings, and full training configurations are provided in Appendix B.

Table 1: Evaluation scores (%). Each column pair reports pass@1 and pass@128 (denoted @1 and @128). Avg is the mean over the five benchmarks. Within each model, **bold** and underlining mark the best and second-best distinct scores in each column.

Method	AIME24		AIME25		AIME26		BeyondAIME		BRUMO25		Avg	
	@1	@128	@1	@128	@1	@128	@1	@128	@1	@128	@1	@128
Qwen2.5-Math-1.5B												
Base	7.29	43.33	3.83	43.33	4.92	36.67	2.00	24.00	7.47	50.00	5.10	39.47
GRPO	11.64	46.67	6.61	40.00	7.60	26.67	2.89	27.00	14.01	43.33	8.55	36.73
R1-zero-Div	7.24	46.67	4.77	46.67	5.57	36.67	2.22	29.00	10.26	43.33	6.01	40.47
DIVER	10.60	50.00	6.98	43.33	7.55	43.33	2.86	30.00	13.46	50.00	8.29	43.33
DQO	11.87	56.67	**7.29**	46.67	**8.88**	40.00	3.37	28.00	14.40	53.33	**9.16**	44.93
G^{2}RL	**12.01**	50.00	6.59	46.67	7.47	40.00	3.20	35.00	14.56	**60.00**	8.77	46.33
GraDiC (ours)	11.17	**63.33**	7.24	**50.00**	7.55	**46.67**	**3.48**	**36.00**	**14.58**	53.33	8.80	**49.87**
Qwen2.5-Math-7B												
Base	11.82	63.33	5.31	53.33	7.27	40.00	2.99	36.00	7.27	56.67	6.93	49.87
GRPO	32.84	73.33	**18.70**	53.33	**14.17**	33.33	**7.37**	34.00	21.93	53.33	19.00	49.46
R1-zero-Div	30.34	66.67	14.51	56.67	11.54	36.67	4.97	41.00	21.17	**63.33**	16.51	52.87
DIVER	33.20	76.67	15.65	53.33	12.08	40.00	5.09	43.00	21.43	56.67	17.49	53.93
DQO	**36.17**	76.67	17.61	53.33	12.58	40.00	6.11	**46.00**	23.31	60.00	**19.16**	55.20
G^{2}RL	33.05	73.33	15.39	**63.33**	12.19	**46.67**	6.11	41.00	**23.37**	56.67	18.02	56.20
GraDiC (ours)	36.15	**80.00**	17.03	**63.33**	12.89	**46.67**	6.09	44.00	22.14	**63.33**	18.86	**59.47**

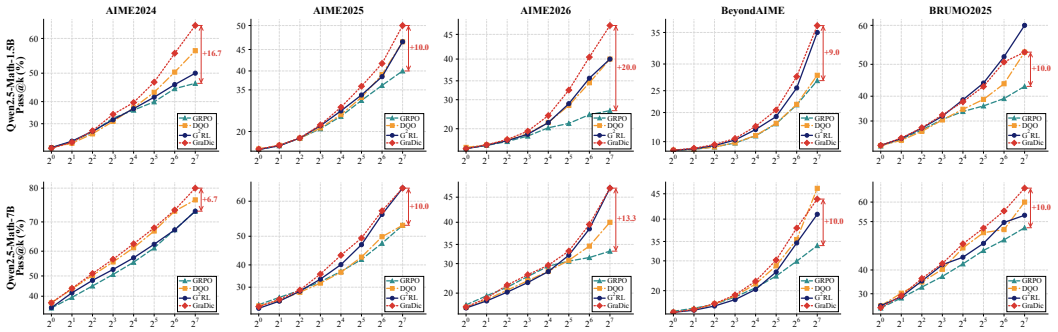


Figure 1: Pass@k performance for $k \in [1, 128]$ across five benchmarks for Qwen2.5-Math-1.5B (top) and Qwen2.5-Math-7B (bottom).

5.2 Main Results

Comparison with baselines. As shown in Table 1, GraDiC maintains competitive pass@1 performance while demonstrating significantly better scaling at larger sampling budgets. At the 1.5B scale, GraDiC achieves an average pass@1 of 8.80%, closely following the highest score of 9.16% from DQO; at the 7B scale, our method reaches 18.86%, comparable to DQO (19.16%) and standard GRPO (19.00%). However, the trends among these methods diverge sharply when evaluated under the diversity metric—pass@128. While standard GRPO sharply improves pass@1 over the base model, the resulting average pass@128 drops from 39.47% to 36.73% (1.5B) and from 49.87% to 49.46% (7B), highlighting the issue of reduced solution coverage after standard post-training. Existing diversity-aware methods manage to recover some of this loss, but GraDiC achieves the highest average pass@128 across both model sizes: 49.87% for the 1.5B model and 59.47% for the 7B model, outperforming the strongest baseline, G^{2}RL (46.33% and 56.20%, respectively).

Pass@k evaluation. Figure 1 shows that the gap between methods generally widens as the sampling budget increases. GraDiC consistently exhibits the strongest scaling trend, widening its

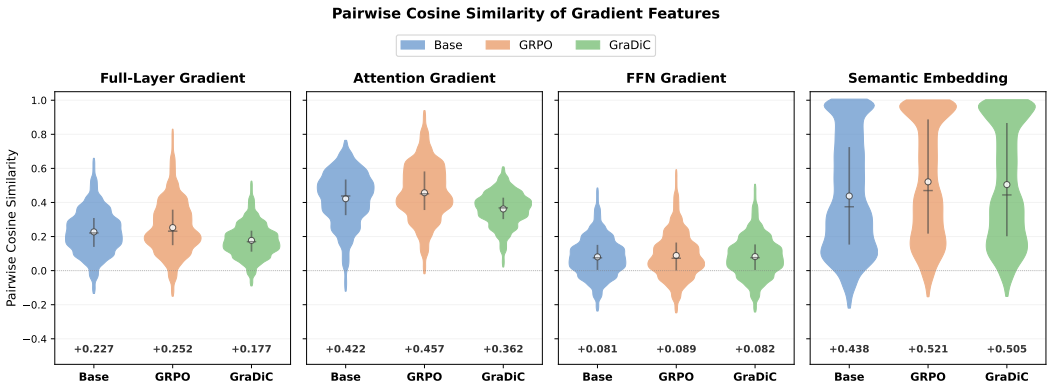


Figure 2: Pairwise cosine similarity of gradient and semantic features for responses sampled from Base, GRPO, and GRADIC Qwen2.5-Math-1.5B on AIME24. Lower values indicate greater diversity in the corresponding feature space.

performance gap over standard GRPO at higher values of k. At $k = 128$, GRADIC improves over GRPO of up to $+20.0\%$ on the 1.5B model, and $+13.3\%$ on the 7B model. The largest absolute improvements over GRPO emerge on AIME26. Given that AIME26 is the most recently released benchmark and strictly post-dates the DAPO-Math-17k training set compared with other benchmarks, this pronounced gap suggests that our gradient-space objective facilitates stronger out-of-distribution (OOD) transfer. BRUMO25 on the 1.5B model is the only setting where G^{2}RL achieves a stronger peak performance than GRADIC at pass@128. Altogether, these results show that our gradient-space objective effectively preserves a broader, more useful distribution of reasoning paths compared to standard post-training.

6 ANALYSIS

Gradient-space diversity after post-training. We conduct a controlled analysis of the response geometry induced by different training methods. For each problem in AIME24, we sample 128 responses from Qwen2.5-Math-1.5B under three settings: base model, standard GRPO, and GRADIC. We measure pairwise cosine similarity between responses using full-layer gradients, gradients from the attention and FFN submodules, and embeddings computed using Qwen3-Embedding-4B [1].

As shown in Figure 2, standard GRPO makes response gradients more similar than those of the base model, with higher similarity for both full-layer and attention gradients. GRPO also exhibits noticeably longer upper tails in these distributions, suggesting the mode collapse issue. In contrast, GRADIC reduces both full-layer and attention-gradient similarity below the base model, producing more diverse gradient directions, while FFN-gradient similarity remains largely unchanged across the three models. Interestingly, semantic similarity follows a different pattern: GRADIC remains more semantically similar than the base model despite having substantially more diverse full-layer gradients. This mismatch indicates that semantic similarity does not necessarily track diversity in the model's update geometry.

7 CONCLUSION

We introduced GRADIC, a principled framework for response-level credit assignment that reweights positive GRPO updates directly in gradient space based on their contribution to an order-2 Vendi objective. This mechanism naturally broadens gradient coverage and mitigates update redundancy while preserving the original reward signal. Empirically, GRADIC demonstrates consistent gains in pass@k on mathematical reasoning benchmarks at a modest training time overhead.

[1] https://huggingface.co/Qwen/Qwen3-Embedding-4B

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and revising the manuscript, checking and revising mathematical proofs, and drafting experimental code. All AI-assisted materials were carefully reviewed and verified by the authors for technical accuracy, correctness, and consistency with the intended message of the paper. In particular, all mathematical proofs were rigorously inspected, and AI-assisted code was tested by the authors before being used in our experiments or included in the final text.

All research ideas, methodological decisions, analyses, interpretations, and contributions presented in this work originated from the authors. Generative AI tools were not used to generate the core research ideas or scientific contributions of this work.

We take full responsibility for the final content of this work, including text, claims, code, mathematical proofs, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work proposes a new training method aimed at improving the capabilities of large language models, thereby contributing to the development of more capable AI systems. Our experiments rely solely on publicly available datasets and open-weight models, and do not involve personally identifiable information (PII), sensitive personal data, or human-subject data. We do not foresee any direct malicious or harmful dual-use implications arising from the proposed method beyond the general risks associated with improving the capabilities of large language models. We conduct and report our experiments with an emphasis on transparency, reproducibility, and responsible research practices.

REPRODUCIBILITY STATEMENT

This work relies solely on publicly available datasets and open-source or open-weight models. We provide detailed descriptions of the proposed method, experimental setup, and key configurations in the main paper and appendix to facilitate reproducibility. Upon publication and subject to internal approval, we commit to release the source code and supporting implementation materials to further enable straightforward reproduction of our results.

REFERENCES

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Creativity and Cognition*, C&C '24, pp. 413–425. ACM, 2024. doi: 10.1145/3635636.3656204. URL http://dx.doi.org/10.1145/3635636.3656204.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1d27c01ebd3e3aebe226b44fc970d803-Abstract-Datasets_and_Benchmarks_Track.html.

ByteDance-Seed. BeyondAIME: Advancing math reasoning evaluation beyond high school olympiads. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME.

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*, 2025. URL https://arxiv.org/abs/2505.12346.

Yilei Chen, Souradip Chakraborty, Lorenz Wolf, Ioannis Paschalidis, and Aldo Pacchiano. Post-training large language models for diverse high-quality responses. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=BV3bLKaJ3Y.

Yuting Deng, Melanie Brucks, and Olivier Toubia. Examining and addressing barriers to diversity in llm-generated ideas. *arXiv preprint arXiv:2602.20408*, 2026. URL https://arxiv.org/abs/2602.20408.

Martin Ehler and Kasso A. Okoudjou. Minimization of the probabilistic p-frame potential. *Journal of Statistical Planning and Inference*, 142(3):645–659, 2012. URL https://arxiv.org/abs/1101.0140.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2023. URL https://arxiv.org/abs/2210.02410.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024. URL https://arxiv.org/abs/2311.09807.

Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Meta-reinforcement learning of structured exploration strategies. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. Diversity-incentivized exploration for versatile reasoning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9G7AbBrd27.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=R0dC7Xzwbk.

Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/4b86abe48d358ecf194c56c69108433e-Paper.pdf.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023. URL https://arxiv.org/abs/2310.06452.

Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2022.03.003. URL http://dx.doi.org/10.1016/j.inffus.2022.03.003.

Zhenwen Liang, Sidi Lu, Wenhao Yu, Kishan Panaganti, Yujun Zhou, Haitao Mi, and Dong Yu. Can llms guide their own exploration? gradient-guided reinforcement learning for llm reasoning. *arXiv preprint arXiv:2512.15687*, 2025. URL https://arxiv.org/abs/2512.15687.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015. URL https://arxiv.org/abs/1504.04909.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Amey P. Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning, 2024. URL https://arxiv.org/abs/2310.12952.

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByBAl2eAZ.

Justin K. Pugh, Lisa B. Soros, and Kenneth O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3, 2016. ISSN 2296-9144. doi: 10.3389/frobt.2016.00040. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2016.00040.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2017. URL https://arxiv.org/abs/1502.05477.

Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, 2010.

Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa Siu, Byron C Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and comparative analysis. In Xuebo Liu and Ayu Purwarianti (eds.), *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 36–46, Mumbai, India, December 2025. Association for Computational Linguistics. ISBN 979-8-89176-301-2. doi: 10.18653/v1/2025.ijcnlp-demo.5. URL https://aclanthology.org/2025.ijcnlp-demo.5/.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of llm generated content. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=O7bF6nlSOD.

Yuda Song, Julia Kempe, and Rémi Munos. Outcome-based exploration for llm reasoning. *arXiv preprint arXiv:2509.06941*, 2025. URL https://arxiv.org/abs/2509.06941.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=PG5fV50maR.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. Diversity-aware policy optimization for large language model reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3169. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8808b4c72d5ade9bbcf9270ac6411314-Abstract-Conference.html.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Suqin Yuan, Jinkun Chen, Jiyang Zheng, Muyang Li, Lei Feng, Dadong Wang, Tao Xiang, Tongliang Liu, and Bo An. Understanding diversity collapse in rlvr via the lens of overtraining. *arXiv preprint arXiv:2606.15455*, 2026. URL https://arxiv.org/abs/2606.15455.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4OsgYD7em5.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*, 2025. URL https://arxiv.org/abs/2509.15194.

A Related Work

Exploration and Diversity in Traditional RL Exploration and diversity have long been central themes in reinforcement learning (Haarnoja et al., 2018). Entropy-regularized methods encourage stochastic policies to avoid premature convergence and maintain broader exploration during training , while quality-diversity approaches such as MAP-Elites explicitly search for repertoires of high-performing yet behaviorally distinct solutions (Mouret & Clune, 2015; Pugh et al., 2016). Related work on unsupervised skill discovery promotes diverse behaviors by maximizing mutual information between latent skills and visited states (Eysenbach et al., 2019). Beyond behavior-level exploration, parameter-space methods perturb policy parameters to induce coherent changes in trajectories (Sehnke et al., 2010; Plappert et al., 2018), and geometry-aware optimization methods such as natural policy gradients (Kakade, 2001) and trust-region approaches (Schulman et al., 2017) structure policy updates through non-Euclidean notions of distance. Collectively, these directions suggest that diversity objectives can be organized at multiple levels, from behaviors and skills to policy parameters and update dynamics, and that maintaining sufficient exploration is central to stable learning and broad generalization.

Evaluating Diversity of LLM Outputs Existing works evaluate LLM output diversity using a range of lexical, syntactic, and embedding-based metrics (Guo et al., 2024; Shaib et al., 2025). Beyond measurement, studies show that post-training procedures like RLHF substantially alter this variation (Kirk et al., 2023; Shypula et al., 2025). This limited diversity ultimately harms downstream utility: it restricts the breadth of ideas in creative tasks compared to human baselines (Deng et al., 2026), and during reasoning tasks, diversity collapse in RLVR narrows solution exploration, i.e., degrading Pass@k performance (Yuan et al., 2026).

Improving Diversity in RL LLMs Recent work has incorporated diversity directly into reinforcement learning objectives for LLMs. For instance, outcome-based exploration (Song et al., 2025) rewards underrepresented answers to preserve solution coverage without sacrificing accuracy, while EVOL-RL (Zhou et al., 2025) balances stable optimization and exploration using novelty-aware rewards. Another line of work promotes exploration by injecting additional stochasticity directly into the RL objective. R1-zero-Div (Yao et al., 2025) introduces a token-level diversity objective that selectively increases entropy on successful trajectories, while DIVER (Hu et al., 2026) moves to sequence-level exploration by constructing rewards from handcrafted structural and lexical diversity measures across sampled solutions. Beyond output stochasticity and heuristics, more recent methods shift to semantic representation spaces. DQO (Chen et al., 2026) applies determinantal point process objectives to similarity kernels constructed from embeddings produced by an external embedding model, while SEED-GRPO (Chen et al., 2025) uses semantic entropy over sampled responses to modulate group-relative policy updates.

GRADIC follows this broader line of work by reshaping within-group credit assignment, but defines its exploration signal directly in the model's optimization space rather than through output-level randomness or external semantic representations. The closest related method, G^2RL (Liang et al., 2025), also uses gradient features, but constructs a heuristic reward from logit-layer gradients whose closed-form dependence on hidden states, which keeps them closely tied to representation space. In contrast, GRADIC operates on full transformer layer gradients and derives per-response credit from a differentiable order-2 Vendi metric, a standard diversity objective, yielding a principled set-level diversity signal with a direct theoretical connection to one-step changes in response probabilities.

A.1 GRPO Notation

For completeness, Table 2 summarizes the notation used in Section 2.1, as well as subsequent sections.

Table 2: Notation for the GRPO objective.

Symbol	Definition
q	input prompt
$o_i = (o_{i,1}, \ldots, o_{i,\lvert o_i \rvert})$	i-th sampled response
G	number of responses sampled for each prompt
$r_i = R(q, o_i)$	verifiable reward for response o_i
$\bar{r}$	mean reward within the sampled group
A_i	group-relative advantage of response o_i
$\pi_{\theta_{\text{old}}}$	behavior policy used to sample responses
π_θ	policy being optimized
$\rho_{i,t}(\theta)$	$\pi_\theta(o_{i,t} \mid q, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})$
ϵ	GRPO clipping parameter
η	learning rate

Table 3: Training and validation hyper-parameters.

Hyperparameter	**Value**
Training batch size	126
# rollouts per prompt	16
PPO mini-batch size	32
Rollout temperature	1.0
Rollout top-p	1.0
Rollout top-k	-1
Validation temperature	1.0
Validation top-p	0.7
Validation top-k	-1
Learning-rate warmup steps	20
Clipping coefficient c	3.0
Lower / upper clipping ratio	0.2 / 0.28
Learning rate	1×10^{-6}
Total epochs	15
Reweighting coefficient β // Vendi objective	0.15
Minimum correct responses // Vendi objective	4
Gradient extraction layer	16

B GRADIC IMPLEMENTATION DETAILS

B.1 TRAINING CONFIGURATIONS

The full algorithm for GRADIC is presented in Algorithm 1. The training and validation hyperparameters are provided in Table 3, while the system prompt is shown in Figure 3. We use gradients at layer 16 for both models to compute the GV_2 reweighting. To ensure stable optimization, we apply the Vendi diversity loss only when a rollout batch achieves at least a 25% success rate; otherwise, we compute the standard GRPO loss. Training and evaluation are conducted on 8 NVIDIA H100 GPUs using VeRL [2] framework.

B.2 EFFICIENT GRADIENT-FEATURE EXTRACTION

In this section, we detail our implementation for efficiently extracting per-rollout gradient features within the VeRL framework, with the complete procedure outlined in Algorithm 2. For any linear module m in the target transformer layer, the per-token contribution to its weight gradient takes the outer-product form $\delta^{(m)}_{\text{out},t}(h^{(m)}_{\text{in},t})^\top$, where $h^{(m)}_{\text{in},t}$ is the input activation and $\delta^{(m)}_{\text{out},t}$ is the loss gradient

[2] https://github.com/verl-project/verl

Algorithm 1 GRADIC reweighting for one GRPO rollout group

Require: Prompt q, rollouts $\{o_i\}_{i=1}^G$, advantages $\{A_i\}_{i=1}^G$, behavior policy π_θ, reweighting strength β, random projection Π

Ensure: Modified advantages $\{\widetilde{A}_i\}_{i=1}^G$

// Identify correct responses and initialize

$\mathcal{P} \leftarrow \{i : A_i > 0\}$; $\widetilde{A}_i \leftarrow A_i$ for all $i \in \{1, \dots, G\}$

// Extract gradient features

for $i \in \mathcal{P}$ **do**

  Compute length-normalized gradient u_i

  $w_i \leftarrow A_i \|u_i\|_2$; project and normalize to obtain unit direction g_i

end for

// Construct the weighted gradient distribution

Normalize masses and form weighted moments:

$$p_i \leftarrow \frac{w_i}{\sum_{j\in\mathcal{P}} w_j}, \qquad K \leftarrow \sum_{i\in\mathcal{P}} p_i g_i g_i^\top, \qquad C \leftarrow \mathrm{tr}(K^2).$$

// Compute GraDiC credits

for $i \in \mathcal{P}$ **do**

  $c_i \leftarrow g_i^\top K g_i$; $D_i \leftarrow 2\left(1 - \frac{c_i}{C}\right)$

end for

// Reweight the positive responses

Apply exponential tilting and rescale advantages:

$$\widetilde{p}_i \propto p_i \exp\big(\beta\ \mathrm{stopgrad}(D_i)\big), \qquad \widetilde{A}_i \leftarrow \frac{\widetilde{p}_i}{p_i} A_i \quad \text{for } i \in \mathcal{P}.$$

return $\{\widetilde{A}_i\}_{i=1}^G$

Algorithm 2 Efficient gradient-feature extraction

Require: Response o, prompt q, actor π_θ, target transformer layer ℓ, hooked linear modules $\mathcal{M}$, response tokens S, random projection Π

Ensure: Unit gradient feature g

// Hook setup and forward pass

Register hooks on $m \in \mathcal{M}$ to record the quantities needed to construct local gradient contributions

Detach the input to layer ℓ as a leaf tensor and run a forward pass on (q, o)

// Truncated backward pass

Form an auxiliary cross-entropy loss L_{aux} over selected response positions S

Compute autograd. $\mathrm{grad}(L_{\mathrm{aux}}, x_\ell)$ *// Backpropagates only down to layer ℓ*

// Project and accumulate gradient features

$z \leftarrow 0$

for $m \in \mathcal{M}$ **do**

  Construct the response gradient contribution $u^{(m)}$ over S

  Project and accumulate: $z \leftarrow z + \mathrm{Proj}_\Pi\left(u^{(m)}\right)$

end for

// Cleanup and normalize

Remove hooks and release the auxiliary computation graph

return $g \leftarrow z/\|z\|_2$

System Prompt

```
Please reason step by step, and put your final answer
within \boxed{}.
```

Figure 3: System prompt used during training and evaluation.

Table 4: End-to-end computational overhead of GRADIC. Runtime is normalized by standard GRPO.

Method	Relative runtime
GRPO	1.000×
GRADIC	1.116×

with respect to the module output. By capturing these two factors via forward and backward hooks, we avoid the prohibitive cost of explicitly materializing a full per-rollout model gradient.

A naive implementation would simply call `loss.backward()` on an auxiliary cross-entropy loss, but this unnecessarily backpropagates through the entire model and pollutes the parameters' `.grad` buffers. To circumvent both overheads, we register a pre-hook on the target layer that replaces its input $\mathbf{h}$ with a detached leaf tensor $\widetilde{\mathbf{h}} = \texttt{detach}(\mathbf{h})\texttt{.requires_grad_(True)}$. Following the forward pass, we evaluate the auxiliary cross-entropy strictly on the selected response positions and invoke $\texttt{torch.autograd.grad}(\mathcal{L}_{\text{aux}}, [\widetilde{\mathbf{h}}])$. This technique successfully propagates gradients through the hooked layer without traversing any preceding layers, while the use of `autograd.grad` ensures that no auxiliary gradients accumulate in the model parameter buffers.

Finally, the captured activation and gradient factors are projected and combined to form the response feature. All hooks and the auxiliary computation graph are then released, ensuring that the subsequent `update_actor` step constructs a fresh graph and performs the standard policy update on the unmodified actor.

Computational Overhead. We first measure the end-to-end training overhead of GRADIC relative to standard GRPO. We restrict this experiment only to Qwen2.5-Math-1.5B due to limited resources. As shown in Table 4, GRADIC increases the total training time by only 11.6% ($1.116\times$). The main additional cost comes from extracting the per-response gradient features, while the subsequent Vendi-based reward computation is negligible in comparison.

To further isolate the dominant overhead, we benchmark the gradient-feature extraction module against a standard forward–backward pass, varying the sequence length from 4k to 20k tokens. As presented in Table 5, once the GPU reaches a steady-throughput regime, the extraction process adds approximately 24% overhead relative to the base operation. This ratio remains remarkably stable as the sequence length grows—ranging from 25.0% at 4k tokens to 23.9% at 20k tokens—because both the gradient-feature extraction and the standard training pass scale approximately linearly with sequence length. This indicates that gradient-feature extraction does not become an increasing computational bottleneck for longer responses.

B.3 Layer Analysis

Layer analysis. We further examine how GRPO updates are distributed across the model. Figure 4(a) reports the relative parameter change $\|\Delta\theta\|/\|\theta\|$ across transformer layers. For both model sizes, the first and last few layers undergo substantially smaller changes, whereas the middle layers exhibit consistently larger updates. Figure 4(b) shows that this trend holds steadily across the entire training process, with middle-layer updates consistently remaining above the overall model average.

One possible interpretation is that RL post-training primarily reshapes intermediate representations involved in composing and processing information for reasoning, while representations closer to the input and output interfaces remain comparatively stable. Although parameter changes alone do not definitively establish this functional interpretation, the concentration of updates in the middle layers suggests they provide a highly informative view of the optimization dynamics induced by RL.

Table 5: Cost of gradient-feature extraction across sequence lengths. "Base" denotes a standard forward–backward operation, and overhead is the extraction time relative to the base operation.

Seq. length	Base	Extraction	Overhead
4000	217.0 ms	54.2 ms	25.0%
8000	462.6 ms	113.9 ms	24.6%
12000	755.6 ms	184.6 ms	24.4%
16000	1085.7 ms	264.6 ms	24.4%
20000	1470.0 ms	351.9 ms	23.9%

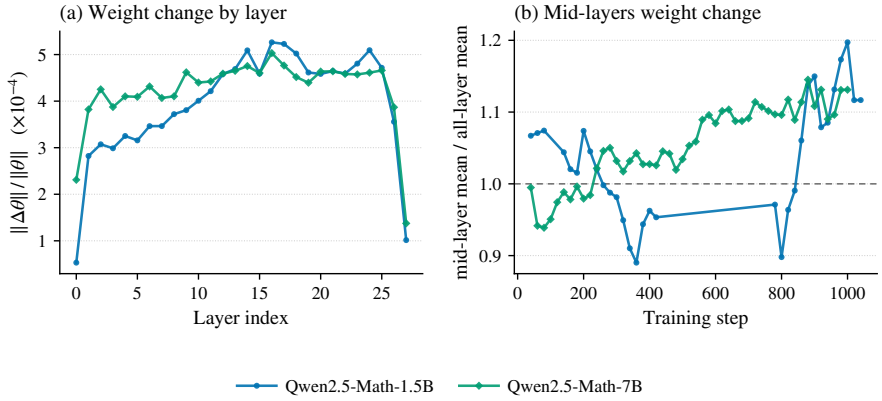


Figure 4: Layer-wise weight changes during training. (a) Relative weight change $\|\Delta\theta\|/\|\theta\|$ across layers. (b) Ratio between the mean weight change of the middle layers and the mean weight change across all layers over training steps.

Figure 5 provides a finer-grained decomposition across parameter modules. Interestingly, the dominant changes in the middle layers arise primarily from the attention projections, whereas the feed-forward modules exhibit substantially smaller updates. This observation aligns with the hypothesis that RL post-training modifies how the model routes and integrates information during generation more strongly than it alters the knowledge encoded in feed-forward parameters. This pattern also motivates future work investigating whether gradients restricted to attention-specific modules are sufficient to characterize, or even control, the learning dynamics of reasoning models.

Taken together, these observations motivate using an intermediate transformer layer as the gradient space for GRADIC. In our setup, we consistently select layer 16 across both models as a representative middle layer. This provides an objective anchor to capture the optimization dynamics of the high-update regime, rather than mining for the single most-changed layer.

C PROOFS AND ADDITIONAL RESULTS

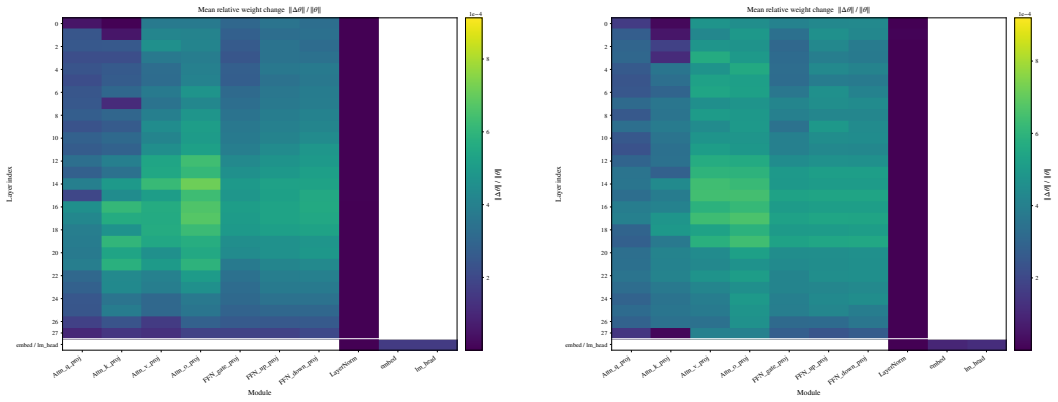
C.1 DETAILED FORMULATION OF SECTION 4.1

Let $\theta_0 = \theta_{\text{old}}$, $\ell_i(\theta) = |o_i|^{-1} \log \pi_\theta(o_i \mid q)$, $u_i = \nabla \ell_i(\theta_0)$, and let U have rows $u_i^\top$. Hold the sampled responses and advantages A fixed. From Section 2.1, the local reward-only GRPO step is $\Delta\theta = \tau U^\top A$, where $\tau = \eta/G$. Assume each ℓ_i is twice continuously differentiable, with $\|\nabla^2 \ell_i(\theta)\|_{\text{op}} \le L_i$ along the update segment. Taylor's theorem gives

$$\Delta\ell_i := \ell_i(\theta_0 + \Delta\theta) - \ell_i(\theta_0) = u_i^\top \Delta\theta + R_i, \qquad |R_i| \le \frac{L_i}{2} \|\Delta\theta\|_2^2. \tag{31}$$

Substituting the update exposes the contribution of each response:

$$\Delta\ell_i = \tau u_i^\top U^\top A + R_i = \tau \sum_{j=1}^{G} \langle u_i, u_j \rangle A_j + R_i. \tag{32}$$



(a) Qwen2.5-Math-1.5B (b) Qwen2.5-Math-7B

Figure 5: Mean relative weight change $\|\Delta\theta\|/\|\theta\|$ across layers and parameter modules for Qwen2.5-Math-1.5B and Qwen2.5-Math-7B. Weight changes are most pronounced in the middle layers, while attention modules exhibit the larger updates compared with feed-forward parameters.

Stacking these identities, with $\Delta\ell = (\Delta\ell_i)_{i=1}^G$ and $R = (R_i)_{i=1}^G$, yields

$$\Delta\ell = \tau UU^\top A + R, \qquad (UU^\top)_{ij} = \langle u_i, u_j \rangle. \tag{33}$$

Since $\|\Delta\theta\|_2^2 = \tau^2 \|U^\top A\|_2^2$, the remainder is second order in τ.

C.2 Proof of the GV_2 Concentration Identity

Proof of Proposition 2. Let $n = |\mathcal{P}|$, $V = \mathrm{span}\{g_1, \ldots, g_n\}$, and $r = \dim V$. Express the gradients in orthonormal coordinates on V. Assume $n \geq 1$, $\|g_i\|_2 = 1$, $p_i \geq 0$, and $\sum_i p_i = 1$. Define

$$X_i = g_i g_i^\top, \qquad K(p) = \sum_i p_i X_i, \qquad C(p) = \mathrm{tr}(K(p)^2), \qquad \mathrm{GV}_2(p) = C(p)^{-1}.$$

Then $\mathrm{tr}(K) = \sum_i p_i \|g_i\|_2^2 = 1$, so

$$\begin{aligned} \left\| K - \frac{I_r}{r} \right\|_F^2 &= \mathrm{tr}\left[\left(K - \frac{I_r}{r} \right)^2 \right] \\ &= \mathrm{tr}(K^2) - \frac{2}{r}\,\mathrm{tr}(K) + \frac{1}{r^2}\,\mathrm{tr}(I_r) \\ &= \mathrm{tr}(K^2) - \frac{1}{r} = \frac{1}{\mathrm{GV}_2(p)} - \frac{1}{r}. \end{aligned} \tag{34}$$

In particular,

$$\mathrm{GV}_2(p) \leq r, \qquad \mathrm{GV}_2(p) = r \iff K(p) = I_r/r.$$

Thus increasing GraDiC's GV_2 objective is equivalent to reducing $\|K(p) - I_r/r\|_F^2$. □

C.3 Proof of the GraDiC Reweighting Result

Derivative of the weighted objective.

Proof of Proposition 1. Let $W_+ = \sum_j w_j$, $p_i = w_i/W_+$, and $K = W_+^{-1} \sum_j w_j X_j$, where $X_i = g_i g_i^\top$. Assume $w_i > 0$ and fixed unit vectors g_i. The quotient rule gives

$$\frac{\partial K}{\partial w_i} = \frac{W_+ X_i - \sum_j w_j X_j}{W_+^2} = \frac{X_i - K}{W_+},$$

$$\frac{\partial\, \mathrm{tr}(K^2)}{\partial w_i} = 2\,\mathrm{tr}\left(K \frac{\partial K}{\partial w_i} \right) = \frac{2}{W_+}(g_i^\top K g_i - C).$$

Consequently,

$$W_+ \frac{\partial \log \mathrm{GV}_2}{\partial w_i} = -\frac{W_+}{C}\frac{\partial C}{\partial w_i} = 2\left(1 - \frac{g_i^\top K g_i}{C}\right) = D_i. \tag{35}$$

Hence $D_i > 0 \iff g_i^\top K g_i < C$ and $D_i = 0 \iff g_i^\top K g_i = C$. GraDiC's credit is therefore proportional to the response's marginal effect on $\log \mathrm{GV}_2$. □

C.3.1 Local ascent

Proof of Theorem 3. Let $X_i = g_i g_i^\top$, $K = \sum_i p_i X_i$, and $C = \mathrm{tr}(K^2)$. Assume $\|g_i\|_2 = 1$, $p_i \geq 0$, $\sum_i p_i = 1$, and fixed gradient vectors during reweighting. Define

$$c_i = g_i^\top K g_i, \qquad D_i = 2\left(1 - \frac{c_i}{C}\right), \qquad Z(\beta) = \sum_j p_j e^{\beta D_j}.$$

Assume the credits D_i are held fixed as β varies. First,

$$\sum_i p_i c_i = \sum_i p_i \,\mathrm{tr}(K X_i) = \mathrm{tr}\left(K \sum_i p_i X_i\right) = C,$$
$$\sum_i p_i D_i = 2\sum_i p_i - \frac{2}{C}\sum_i p_i c_i = 0.$$

For GraDiC's weights $p_i(\beta) = p_i e^{\beta D_i}/Z(\beta)$, we have $Z(0) = 1$ and $Z'(0) = \sum_i p_i D_i = 0$. Therefore

$$p_i'(0) = \left.\frac{p_i D_i e^{\beta D_i} Z(\beta) - p_i e^{\beta D_i} Z'(\beta)}{Z(\beta)^2}\right|_{\beta=0} = p_i D_i, \tag{36}$$
$$\left.\frac{dK(p(\beta))}{d\beta}\right|_0 = \sum_i p_i D_i X_i.$$

Differentiating the objective gives

$$\begin{aligned}\left.\frac{d}{d\beta}\log \mathrm{GV}_2(p(\beta))\right|_0 &= -\frac{2}{C}\,\mathrm{tr}\left(K\sum_i p_i D_i X_i\right)\\ &= -\frac{2}{C}\sum_i p_i D_i c_i\\ &= \sum_i p_i D_i (D_i - 2)\\ &= \sum_i p_i D_i^2 = \mathrm{Var}_{i\sim p}(D_i) \geq 0.\end{aligned} \tag{37}$$

The equality condition is

$$\mathrm{Var}_{i\sim p}(D_i) = 0 \quad \iff \quad D_i = 0 \quad \text{for every } i \text{ with } p_i > 0.$$

Otherwise, continuity of the derivative gives

$$\exists \beta_0 > 0: \qquad \mathrm{GV}_2(p(\beta)) > \mathrm{GV}_2(p) \quad \text{for all } 0 < \beta < \beta_0.$$

Under these assumptions, GraDiC's reweighting therefore increases the rollout group's weighted gradient diversity for sufficiently small positive steps whenever the variance above is positive. □

C.3.2 Finite-step improvement

The local-ascent theorem gives the change in gradient diversity at $\beta = 0$. The result below gives a sufficient condition for a finite reweighting step to increase GV_2, both when using GraDiC's credits directly and when applying a nondecreasing transformation to them before exponential reweighting.

Theorem 5 (Finite-step improvement with transformed credits)**.** *Assume $p_i > 0$, $\sum_i p_i = 1$, fixed unit gradients, and $S_i = \psi(D_i)$, where ψ is nondecreasing and L_ψ-Lipschitz with $L_\psi > 0$. For $\beta > 0$, define*

$$q_i = \frac{p_i e^{\beta S_i}}{\sum_j p_j e^{\beta S_j}}, \qquad J(p,q) = \sum_i (q_i - p_i) \log \frac{q_i}{p_i}.$$

Then

$$C(q) \le C(p) - \left(\frac{C(p)}{\beta L_\psi} - \frac{1}{2} \right) J(p,q). \tag{38}$$

Proof of Theorem 5. Write $K = K(p), C = C(p), \delta = q - p$, and $\Delta K = K(q) - K(p) = \sum_i \delta_i X_i$. Since $\sum_i \delta_i = 0$,

$$\begin{aligned} C(q) - C &= 2 \operatorname{tr}(K \Delta K) + \|\Delta K\|_F^2 \\ &= 2 \sum_i \delta_i c_i + \|\Delta K\|_F^2 \\ &= -C \sum_i \delta_i D_i + \|\Delta K\|_F^2. \end{aligned} \tag{39}$$

For $q \neq p$, set $t = \|\delta\|_1 / 2$ and define

$$M_+ = \frac{1}{t} \sum_{\delta_i > 0} \delta_i X_i, \qquad M_- = \frac{1}{t} \sum_{\delta_i < 0} (-\delta_i) X_i.$$

Both matrices are positive semidefinite with trace one. Therefore

$$\begin{aligned} \|\Delta K\|_F^2 &= t^2 \operatorname{tr}[(M_+ - M_-)^2] \\ &= t^2 \big[\operatorname{tr}(M_+^2) + \operatorname{tr}(M_-^2) - 2 \operatorname{tr}(M_+ M_-) \big] \\ &\le 2t^2 = \frac{1}{2} \|q - p\|_1^2. \end{aligned} \tag{40}$$

For $a, b > 0$, $(a - b) \log(a/b) \ge 2(a-b)^2/(a+b)$, with equality if and only if $a = b$. To see this, for $x \ge 1$ use

$$\log x - \frac{2(x-1)}{x+1} = \int_1^x \frac{(t-1)^2}{t(t+1)^2} \, dt \ge 0,$$

and interchange a, b when $a < b$. Applying this inequality and Cauchy–Schwarz yields

$$\begin{aligned} J(p,q) &= \mathrm{KL}(p \| q) + \mathrm{KL}(q \| p) \\ &\ge 2 \sum_i \frac{(q_i - p_i)^2}{p_i + q_i} \\ &\ge \frac{2(\sum_i |q_i - p_i|)^2}{\sum_i (p_i + q_i)} = \|q - p\|_1^2. \end{aligned} \tag{41}$$

In particular, $\|\Delta K\|_F^2 \le J(p,q)/2$, with strict inequality when $q \neq p$.

Next put $Z = \sum_i p_i e^{\beta S_i}$. For any numbers a_i,

$$\begin{aligned} \sum_i (q_i - p_i) a_i &= \frac{1}{Z} \sum_{i,j} p_i p_j a_i (e^{\beta S_i} - e^{\beta S_j}) \\ &= \frac{1}{2Z} \sum_{i,j} p_i p_j (a_i - a_j)(e^{\beta S_i} - e^{\beta S_j}). \end{aligned} \tag{42}$$

The assumptions on ψ imply

$$(D_i - D_j)(S_i - S_j) \ge 0, \qquad |S_i - S_j| \le L_\psi |D_i - D_j|,$$

and hence

$$(D_i - D_j)(e^{\beta S_i} - e^{\beta S_j}) \ge \frac{1}{L_\psi} (S_i - S_j)(e^{\beta S_i} - e^{\beta S_j}). \tag{43}$$

Using Eq. (42) with $a_i = D_i$ and $a_i = S_i$, respectively, gives

$$\sum_i \delta_i D_i \geq \frac{1}{L_\psi} \sum_i \delta_i S_i$$
$$= \frac{1}{\beta L_\psi} \sum_i \delta_i \left(\log \frac{q_i}{p_i} + \log Z \right) = \frac{J(p,q)}{\beta L_\psi}. \tag{44}$$

Substituting into Eq. (39),

$$C(q) - C(p) \leq -\frac{C(p)}{\beta L_\psi} J(p,q) + \frac{1}{2} J(p,q).$$

Equality in Eq. (38) holds if and only if

$$q = p \quad \Longleftrightarrow \quad S_i = S_j \quad \text{for all } i, j.$$

It follows that

$$0 < \beta L_\psi < 2C(p), \quad q \neq p \quad \Longrightarrow \quad \mathrm{GV}_2(q) > \mathrm{GV}_2(p).$$

□

For GraDiC's original credit, $\psi(x) = x$ and $L_\psi = 1$, so $0 < \beta < 2C(p)$ guarantees an increase in GV_2 whenever the weights change. The quantity $J(p,q)$ measures the change from the initial mass distribution p to the reweighted distribution q. Multiplying it by the positive coefficient in Eq. (38) bounds the resulting decrease in inverse diversity C from below. This extends the first-order guarantee to a nonzero step size, with the allowable step depending on the group's current concentration and the scale of the credit transformation.

C.4 Proof of the OOD Transfer Result

C.4.1 Transfer under Orthogonal Modes

Proof of Theorem 4. Let $e_1, \ldots, e_r$ denote gradient modes, and let $\rho_c = \sum_{i:g_i = e_c} p_i$. Assume

$$e_c^\top e_{c'} = \mathbf{1}_{\{c=c'\}}, \qquad g_i \in \{e_1, \ldots, e_r\}, \qquad p_i \geq 0, \qquad \sum_i p_i = 1.$$

Then $\sum_c \rho_c = 1$ and

$$K = \sum_c \rho_c e_c e_c^\top,$$
$$\mathrm{tr}(K^2) = \sum_{c,c'} \rho_c \rho_{c'} (e_c^\top e_{c'})^2 = \sum_c \rho_c^2 = \frac{1}{\mathrm{GV}_2(\rho)}.$$

For $\bar{\xi} \in \mathbb{R}$ and $\sigma \geq 0$, assume the transfer coefficients belong to

$$\mathcal{U}(\bar{\xi}, \sigma) = \left\{ \xi \in \mathbb{R}^r : \frac{1}{r} \sum_c \xi_c = \bar{\xi}, \quad \frac{1}{r} \sum_c (\xi_c - \bar{\xi})^2 \leq \sigma^2 \right\}. \tag{45}$$

Put $u_r = \mathbf{1}/r$ and $z = \xi - \bar{\xi}\mathbf{1}$. The constraints give $\mathbf{1}^\top z = 0$ and $\|z\|_2 \leq \sqrt{r}\,\sigma$. Therefore

$$\rho^\top \xi = \bar{\xi} + \rho^\top z$$
$$= \bar{\xi} + (\rho - u_r)^\top z$$
$$\geq \bar{\xi} - \|\rho - u_r\|_2 \|z\|_2$$
$$\geq \bar{\xi} - \sqrt{r}\,\sigma \|\rho - u_r\|_2. \tag{46}$$

Also,

$$\|\rho - u_r\|_2^2 = \sum_c \rho_c^2 - \frac{2}{r} \sum_c \rho_c + \frac{1}{r}$$
$$= \sum_c \rho_c^2 - \frac{1}{r} = \frac{1}{\mathrm{GV}_2(\rho)} - \frac{1}{r}.$$

Equality is attained by

$$\xi^\star = \begin{cases} \bar{\xi}\mathbf{1} - \sqrt{r}\,\sigma \dfrac{\rho - u_r}{\|\rho - u_r\|_2}, & \rho \neq u_r, \\ \bar{\xi}\mathbf{1}, & \rho = u_r. \end{cases} \tag{47}$$

For $\rho \neq u_r$, this vector satisfies $\mathbf{1}^\top(\xi^\star - \bar{\xi}\mathbf{1}) = 0$ and $\|\xi^\star - \bar{\xi}\mathbf{1}\|_2 = \sqrt{r}\,\sigma$. For $\rho = u_r$, every $\xi \in \mathcal{U}$ gives $\rho^\top \xi = \bar{\xi}$. Consequently,

$$\inf_{\xi \in \mathcal{U}(\bar{\xi},\sigma)} \rho^\top \xi = \bar{\xi} - \sigma\sqrt{\frac{r}{\mathrm{GV}_2(\rho)} - 1}.$$

Under the stated mode assumptions, increasing GV_2 raises this worst-case transfer bound for fixed $\bar{\xi}$ and σ. □

C.4.2 Transfer under Approximate Modes

We extend the transfer result to groups whose gradients lie near orthogonal modes, allowing responses assigned to the same mode to have different directions. The bound quantifies how the average distance to these modes changes the relationship between measured gradient diversity and guaranteed target alignment.

Proof of the approximate-mode transfer bound. Let $e_1, \ldots, e_m$ be orthonormal vectors and let $a(i)$ assign each g_i to a mode. Define

$$\begin{aligned} \rho_c &= \sum_{i:a(i)=c} p_i, \qquad \varepsilon = \sum_i p_i \|g_i - e_{a(i)}\|_2, \qquad \bar{g} = \sum_i p_i g_i, \\ K &= \sum_i p_i g_i g_i^\top, \qquad K_0 = \sum_c \rho_c e_c e_c^\top, \qquad C = \|K\|_F^2. \end{aligned} \tag{48}$$

Assume $\|g_i\|_2 = 1$, $p_i \geq 0$, $\sum_i p_i = 1$, $\|h_T\|_2 \leq H_T$, and $\xi = (h_T^\top e_c)_{c=1}^m \in \mathcal{U}(\bar{\xi}, \sigma)$. Writing $\delta = \bar{g} - \sum_c \rho_c e_c$, we obtain

$$\begin{aligned} \|\delta\|_2 &= \left\| \sum_i p_i (g_i - e_{a(i)}) \right\|_2 \leq \sum_i p_i \|g_i - e_{a(i)}\|_2 = \varepsilon, \\ \|g_i g_i^\top - e_{a(i)} e_{a(i)}^\top\|_F &= \|(g_i - e_{a(i)}) g_i^\top + e_{a(i)}(g_i - e_{a(i)})^\top\|_F \\ &\leq 2\|g_i - e_{a(i)}\|_2, \\ \|K - K_0\|_F &\leq 2 \sum_i p_i \|g_i - e_{a(i)}\|_2 = 2\varepsilon. \end{aligned}$$

Since $\|K\|_F \leq 1$, $\|K_0\|_F \leq 1$, and $\|K_0\|_F^2 = \sum_c \rho_c^2$, it follows that

$$\begin{aligned} \left| C - \sum_c \rho_c^2 \right| &= |\operatorname{tr}[(K - K_0)(K + K_0)]| \\ &\leq \|K - K_0\|_F \|K + K_0\|_F \leq 4\varepsilon. \end{aligned} \tag{49}$$

Thus $1/m \leq \sum_c \rho_c^2 \leq \min\{1, C + 4\varepsilon\}$. Applying the orthogonal-mode bound to ρ gives

$$\begin{aligned} h_T^\top \bar{g} &= \rho^\top \xi + h_T^\top \delta \\ &\geq \bar{\xi} - \sigma\sqrt{m \sum_c \rho_c^2 - 1} - H_T \varepsilon \\ &\geq \bar{\xi} - \sigma\sqrt{m \min\{1, C + 4\varepsilon\} - 1} - H_T \varepsilon. \end{aligned} \tag{50}$$

Equality in the final bound holds precisely when

$$\rho^\top \xi = \bar{\xi} - \sigma\sqrt{m \sum_c \rho_c^2 - 1}, \qquad h_T^\top \delta = -H_T \varepsilon,$$

$$\sigma \left[\sqrt{m \min\{1, C + 4\varepsilon\} - 1} - \sqrt{m \sum_c \rho_c^2 - 1} \right] = 0.$$

□

Thus, the transfer interpretation of GV_2 remains robust even when gradient modes are only approximate. For a fixed mode error and target assumptions, increasing $\mathrm{GV}_2(p)$ will consistently improve or maintain the transfer lower bound. The approximation simply introduces two controlled error terms: 4ε captures the mismatch between $C = 1/\,\mathrm{GV}_2(p)$ and the actual mode-mass concentration, while $H_T\varepsilon$ accounts for the resulting alignment error. Naturally, when the modes are perfectly orthogonal ($\varepsilon = 0$), this bound smoothly recovers the exact result.

C.5 Additional Properties of the Vendi Objective

C.5.1 Coverage Optimality and Stationary Conditions

This result identifies the largest diversity attainable by redistributing mass among a fixed set of rollout gradients. It also relates GraDiC's response credits to the remaining gap from that optimum, giving a criterion for when further redistribution cannot improve the objective.

Corollary 6 (Optimality on a fixed set of directions)**.** *For fixed unit gradients, let* $\Delta_n = \{q_i \geq 0 : \sum_i q_i = 1\}$ *and* $C_* := \min_{q\in\Delta_n} C(q)$*. For any* $p \in \Delta_n$*,*

$$0 \leq C(p) - C_* \leq C(p) \max_i D_i(p). \tag{51}$$

Furthermore,

$$C(p) = C_* \quad \Longleftrightarrow \quad D_i(p) \leq 0 \text{ and } p_i D_i(p) = 0 \quad \text{for every } i.$$

Proof of Corollary 6. Let $X_i = g_i g_i^\top$, $K(p) = \sum_i p_i X_i$, and $\mathcal{K} = \mathrm{conv}\{X_1, \ldots, X_n\}$. Assume $\|g_i\|_2 = 1$ and write $r = \dim \mathrm{span}\{g_i\}$. The minimum is attained because $\mathcal{K}$ is compact. Choose a minimizing distribution q_* and let $K_* = K(q_*)$. The quadratic expansion gives

$$\begin{aligned} C_* - C(p) &= 2\sum_i (q_{*,i} - p_i) c_i(p) + \|K_* - K(p)\|_F^2 \\ &= -C(p) \sum_i q_{*,i} D_i(p) + \|K_* - K(p)\|_F^2. \end{aligned}$$

Consequently,

$$\begin{aligned} C(p) - C_* &= C(p) \sum_i q_{*,i} D_i(p) - \|K_* - K(p)\|_F^2 \\ &\leq C(p) \sum_i q_{*,i} D_i(p) \leq C(p) \max_i D_i(p). \end{aligned}$$

For necessity of the optimality condition, let b_i be the ith coordinate vector of $\mathbb{R}^n$. At a minimizer p,

$$\frac{d}{dt} C((1-t)p + t b_i)\Big|_{t=0^+} = 2(c_i(p) - C(p)) = -C(p) D_i(p) \geq 0.$$

Thus $D_i(p) \leq 0$ for every i. Since $\sum_i p_i D_i(p) = 0$, also $p_i D_i(p) = 0$ for every i. Conversely, these conditions imply that, for any $q \in \Delta_n$,

$$C(q) - C(p) = -C(p) \sum_i q_i D_i(p) + \|K(q) - K(p)\|_F^2 \geq 0.$$

Both bounds in Eq. (51) are equalities exactly when $C(p) = C_*$. For positive weights, the condition reduces to

$$p_i > 0 \; (1 \leq i \leq n): \qquad C(p) = C_* \quad \Longleftrightarrow \quad D_i(p) = 0 \; (1 \leq i \leq n).$$

Finally, the concentration identity gives the attainable maximum

$$\max_{q\in\Delta_n} \mathrm{GV}_2(q) = \frac{1}{C_*} \leq r, \qquad C_* = \frac{1}{r} \quad \Longleftrightarrow \quad I_r/r \in \mathcal{K}. \tag{52}$$

The minimizing matrix is unique: if K_1, K_2 both minimize the squared Frobenius norm, then

$$\left\| \frac{K_1 + K_2}{2} \right\|_F^2 = \frac{\|K_1\|_F^2 + \|K_2\|_F^2}{2} - \frac{1}{4}\|K_1 - K_2\|_F^2 = C_* - \frac{1}{4}\|K_1 - K_2\|_F^2 \geq C_*,$$

which forces $K_1 = K_2$. Equation (52) is the weighted directional-coverage optimum (Ehler & Okoudjou, 2012). □

Ultimately, GRADIC's credits offer an intuitive measure of optimality: a positive credit $D_i > 0$ highlights a response whose weight should be increased to improve gradient coverage, while a state where $D_i = 0$ for all active responses signals an optimal allocation. Furthermore, the largest positive credit provides a practical bound on the remaining suboptimality, given by $C(p) - C_* \leq C(p) \max_i D_i(p)$. This optimal coverage reaches its theoretical maximum of $\mathrm{GV}_2 = r$ exactly when the gradient directions can be perfectly balanced into an isotropic weighting, $K = I_r/r$.

Influence of an additional direction. GraDiC's credit can also be interpreted by asking how diversity changes when a small fraction of the group's mass is assigned to a candidate gradient direction. The calculation below applies even when that direction is not already among the sampled responses.

Proof of the derivative for an additional direction. Let K be the current second moment and $C = \mathrm{tr}(K^2)$. Assume g is a fixed unit vector and define $K_\varepsilon = (1-\varepsilon)K + \varepsilon g g^\top$ for $0 \leq \varepsilon \leq 1$. Then

$$\begin{aligned} \frac{d}{d\varepsilon} \log \frac{1}{\mathrm{tr}(K_\varepsilon^2)}\bigg|_{0^+} &= -\frac{2}{C}\,\mathrm{tr}[K(gg^\top - K)] \\ &= 2\left(1 - \frac{g^\top K g}{C}\right). \end{aligned} \tag{53}$$

The derivative is positive if $g^\top K g < C$ and zero if $g^\top K g = C$. □

The overlap $g^\top K g$ compares the candidate direction with the current weighted gradient group, while C is the group's average squared overlap. Assigning mass to a direction with overlap below this average increases diversity locally; assigning it to a direction above the average decreases diversity. Setting $g = g_i$ recovers GraDiC's credit D_i, so the response credit measures the benefit of shifting mass toward that direction from the current mixture.

C.5.2 QUADRATIC-MMD IDENTITIES

We connect the gradient-diversity objective to maximum mean discrepancy (MMD), a discrepancy between distributions (Gretton et al., 2012). With the squared-inner-product kernel, maximizing GV_2 is equivalent to reducing MMD between the weighted empirical gradient distribution and an isotropic reference distribution.

Proof of the quadratic-MMD identities. Let $P_p = \sum_i p_i \delta_{g_i}$, $K = \sum_i p_i g_i g_i^\top$, and $\kappa(g,h) = \langle g, h\rangle^2$. Assume g_i are unit vectors in an r-dimensional space, $p_i \geq 0$, and $\sum_i p_i = 1$. Let Q be a distribution on the same unit sphere and define $K_Q = \mathbb{E}_{h\sim Q}[hh^\top]$. Assume distinct random vectors in each expectation below are drawn independently. The kernel is positive semidefinite, as for any real a_i, we have:

$$\sum_{i,j} a_i a_j \kappa(g_i, g_j) = \left\| \sum_i a_i g_i g_i^\top \right\|_F^2 \geq 0.$$

Expanding the squared MMD gives

$$\begin{aligned} \mathrm{MMD}_\kappa^2(P_p, Q) &= \mathbb{E}_{g,g'\sim P_p}\kappa(g,g') - 2\mathbb{E}_{g\sim P_p, h\sim Q}\kappa(g,h) + \mathbb{E}_{h,h'\sim Q}\kappa(h,h') \\ &= \mathrm{tr}(K^2) - 2\,\mathrm{tr}(KK_Q) + \mathrm{tr}(K_Q^2) \\ &= \|K - K_Q\|_F^2. \end{aligned} \tag{54}$$

Assume now $K_Q = I_r/r$. Since $\mathrm{tr}(K) = 1$,

$$\mathrm{MMD}_\kappa^2(P_p, Q) = \left\| K - \frac{I_r}{r} \right\|_F^2 = \frac{1}{\mathrm{GV}_2(p)} - \frac{1}{r}. \tag{55}$$

For symmetric H with $\|H\|_F \leq 1$, the quadratic-test difference is

$$\begin{aligned} &\mathbb{E}_{P_p}[g^\top H g] - \mathbb{E}_Q[h^\top H h] = \mathrm{tr}[H(K - I_r/r)], \\ &\quad |\,\mathrm{tr}[H(K - I_r/r)]| \leq \|H\|_F \|K - I_r/r\|_F \leq \|K - I_r/r\|_F. \end{aligned}$$

The supremum equals $\|K - I_r/r\|_F$:

$$\sup_{H=H^\top,\, \|H\|_F \le 1} \left|\mathbb{E}_{P_p}[g^\top H g] - \mathbb{E}_Q[h^\top H h]\right| = \|K - I_r/r\|_F,$$

with equality attained by

$$H^\star = \begin{cases} (K - I_r/r)/\|K - I_r/r\|_F, & K \neq I_r/r, \\ 0, & K = I_r/r. \end{cases}$$

Finally, define $f_p(g) = \mathbb{E}_{h\sim P_p}\kappa(g,h) - \mathbb{E}_{h\sim Q}\kappa(g,h)$. Then

$$\begin{aligned} f_p(g) &= g^\top K g - \frac{1}{r}, \\ \mathbb{E}_{P_p} f_p &= C - \frac{1}{r}, \\ -\frac{2}{C}\big(f_p(g_i) - \mathbb{E}_{P_p} f_p\big) &= -\frac{2}{C}(c_i - C) = D_i. \end{aligned} \tag{56}$$

In particular,

$$\mathrm{MMD}^2_\kappa(P_p, Q) = 0 \quad\Longleftrightarrow\quad K = I_r/r \quad\Longleftrightarrow\quad \mathrm{GV}_2(p) = r.$$

□

Thus, increasing GV_2 moves the weighted empirical distribution closer to a reference with second moment I_r/r in MMD, achieving a more even distribution of second-moment mass across directions. Furthermore, Equation 56 also connects this distribution-level objective to individual credits: $f_p(g_i)$ measures how much more response i overlaps with the current group than with the reference. Because $D_i > 0$ when this excess overlap is below average, GRADIC's reweighting therefore favors directions that reduce this discrepancy.

D ORDER-1 VENDI CREDIT ASSIGNMENT

The weight-differentiation construction used by GraDiC also gives response-level credits for the order-1 Vendi Score. We derive these credits by replacing the order-2 objective with the entropy of the gradient second moment's eigenvalues, while retaining the same normalized response weights.

Derivation of the order-1 response credit. Let $X_i = g_i g_i^\top$, $W_+ = \sum_i w_i$, $p_i = w_i/W_+$, and $K = \sum_i p_i X_i$. Assume fixed unit vectors g_i, positive weights w_i, and $K \succ 0$ in orthonormal coordinates on $\mathrm{span}\{g_i\}$. Define

$$\Phi_1(w) = \log \mathrm{GV}_1(w) = -\operatorname{tr}(K \log K). \tag{57}$$

For $f(x) = x \log x$, the trace derivative is

$$d \operatorname{tr} f(K) = \operatorname{tr}(f'(K)\, dK) = \operatorname{tr}[(\log K + I) dK].$$

Differentiating the normalized second moment gives

$$\begin{aligned} \frac{\partial K}{\partial w_i} &= \frac{W_+ X_i - \sum_j w_j X_j}{W_+^2} = \frac{X_i - K}{W_+}, \\ \operatorname{tr}(X_i - K) &= \|g_i\|_2^2 - \operatorname{tr}(K) = 0. \end{aligned}$$

The response credit is therefore

$$\begin{aligned} D_i^{(1)} := W_+ \frac{\partial \Phi_1}{\partial w_i} &= -\operatorname{tr}[(\log K + I)(X_i - K)] \\ &= -\operatorname{tr}[(\log K)(X_i - K)] \\ &= \operatorname{tr}(K \log K) - g_i^\top (\log K) g_i. \end{aligned} \tag{58}$$

Its weighted mean satisfies

$$\sum_i p_i D_i^{(1)} = \operatorname{tr}(K \log K) - \operatorname{tr}\left[\left(\sum_i p_i X_i\right) \log K\right] = 0. \tag{59}$$

□

The credit $D_i^{(1)}$ measures the marginal increase in $\log \mathrm{GV}_1$ from assigning more mass to response i. Its zero weighted mean makes the comparison relative to the current gradient mixture, as for the order-2 credit. Substituting $D_i^{(1)}$ into GraDiC's exponential reweighting therefore gives an order-1 version of the same credit-assignment mechanism; its local-ascent property follows by the same differentiation argument as Appendix C.3.1.

E Additional Experimental Results

GV_2 and GV_1 Performance Comparison Table 6 demonstrates that GV_1 achieves slightly higher average pass@1 performance than GV_2 (9.19 vs. 8.80), whereas GV_2 performs better at larger sampling budgets, improving the average pass@128 from 48.13 to 49.87.

Table 6: Ablation study of the Vendi order on Qwen2.5-Math-1.5B. Evaluation scores are reported in %. Each column pair reports pass@1 and pass@128 (denoted @1 and @128). Avg is the mean over the five benchmarks. **Bold** and underlining mark the best and second-best distinct scores in each column, respectively.

Method	AIME24		AIME25		AIME26		BeyondAIME		BRUMO25		Avg	
	@1	@128	@1	@128	@1	@128	@1	@128	@1	@128	@1	@128
GraDiC (GV_2)	<u>11.17</u>	**63.33**	<u>7.24</u>	<u>50.00</u>	**7.55**	**46.67**	<u>3.48</u>	**36.00**	**14.58**	**53.33**	<u>8.80</u>	**49.87**
GraDiC (GV_1)	**11.72**	<u>60.00</u>	**8.72**	**53.33**	<u>7.45</u>	<u>43.33</u>	**3.52**	<u>34.00</u>	<u>14.53</u>	<u>50.00</u>	**9.19**	<u>48.13</u>

Layer choice. We run additional experiments using gradients extracted from different layers for GraDiC. Table 7 demonstrates that the last layer achieves the highest average pass@1, while Layer 16 yields competitive pass@1 and the highest average pass@128. We additionally consider gradients of the logit layer, as used in Liang et al. (2025). Let h_t denote the final hidden state, W the output projection (logit layer), and $p_t = \mathrm{softmax}(Wh_t)$. The gradient of the log-probability of token o_t with respect to W admits the closed-form expression:

$$\nabla_W \log p_t(o_t) = (e_{o_t} - p_t)\, h_t^\top,$$

and the response-level gradient is:

$$u_i^{\mathrm{logit}} = \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left(e_{o_{i,t}} - p_{i,t}\right) h_{i,t}^\top.$$

Despite its computational convenience, the logit-layer gradient performs worst on both average pass@1 and pass@128. This supports our argument that gradients from a full transformer layer provide a more informative training signal than output-layer gradients.

Table 7: Ablation study of the gradient extraction layer on Qwen2.5-Math-1.5B. Evaluation scores are reported in %. We use Minerva-Math[4] as the held-out validation set.

Gradient Layer	Pass@1	Pass@128
Logit Layer	20.03	58.46
Last Transformer Layer	**21.08**	<u>61.4</u>
Layer 16	<u>20.73</u>	**62.59**

Reweighting Sensitivity We further investigate the effect of the reweighting strength β in Eq. 21. Recall that β controls how strongly the Vendi credit modifies the relative update. Consider two responses with orthogonal gradient directions and initial masses $p_1 = 0.25$ and $p_2 = 0.75$. This

[4] https://huggingface.co/datasets/math-ai/minervamath

Table 8: Ablation study of the reweighting strength β on Qwen2.5-Math-1.5B. Due to computational constraints, we train the model on a random subset of 2,000 training samples (out of the full 17k dataset) for 500 steps. We use Minerva-Math as the held-out validation set.

β	Pass@1	Pass@128
0.05	**18.10**	51.03
0.10	<u>17.61</u>	53.48
0.15	17.58	**57.73**
0.25	17.11	<u>56.90</u>
0.50	15.82	54.58
1.00	14.40	48.79

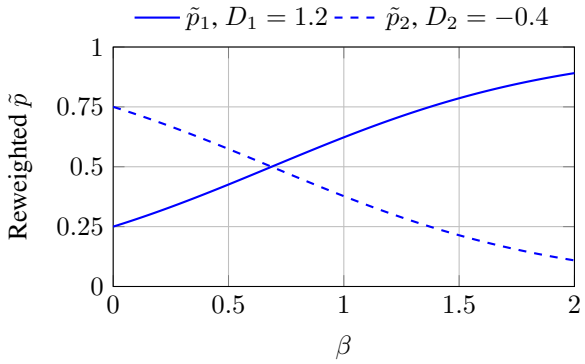


Figure 6: Toy illustration of the effect of the reweighting strength β. Starting from $p_1 = 0.25$ and $p_2 = 0.75$, increasing β shifts update mass toward the underrepresented response.

gives $c_1 = 0.25$, $c_2 = 0.75$, and $C = 0.625$, yielding valid Vendi credits $D_1 = 1.2$ and $D_2 = -0.4$. Their relative mass after reweighting becomes

$$\frac{\tilde{p}_1}{\tilde{p}_2} = \frac{1}{3} e^{1.6\beta}.$$

Figure 6 plots the resulting mass distributions as a function of β. When $\beta = 0$, the original mass distribution remains unchanged. As β increases, the exponential reweighting progressively shifts more positive update mass toward the higher-credit, underrepresented response.

Table 8 shows the effect of varying the reweighting strength β. Small values, such as $\beta = 0.05$, produce only mild changes to the standard GRPO baseline ($\beta = 0$) and maintain moderate performance. Values in the range $\beta \in [0.1, 0.25]$ perform consistently well, with $\beta = 0.15$ achieving the highest pass@128. In contrast, overly strong reweighting, such as $\beta = 1$, leads to unstable training and degrades final performance.