UNDERSTANDING AND EXPLOITING ANISOTROPY IN POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM post-training combines supervised fine-tuning (SFT), a mode-covering forward-KL objective, with reinforcement learning (RL), a mode-seeking reverse-KL objective. Frequency-weighted likelihood training leaves a well-known signature: *anisotropy*, in which a few residual channels carry disproportionately large activations. Anisotropy is widely documented and usually treated as a defect, yet its function and its interaction with post-training remain unclear. We first analyze it. A label-free outlier rule isolates about 5% of residual channels that are essential for language modeling: removing them raises perplexity from 10 to over 10^6, versus 35 for count-matched random channels. Yet they barely distinguish correct from incorrect reasoning. SFT reshapes them, whereas RL leaves them largely intact and adapts the complementary channels. These channels therefore form the model's *coherence substrate*, and reasoning adaptation happens elsewhere. We then exploit this. SPHEREGATE learns one bounded gain per residual channel on a frozen backbone. Its activation-weighted gradients provably limit movement of high-energy coherence channels and leave the remaining channels free. With 0.1M trainable parameters, SPHEREGATE outperforms parameter-efficient baselines by 2.0–7.3 points on MATH-500 across Qwen2.5 (0.5B–7B) and Llama-3-8B, is comparable or exceeds full-model GRPO. Anisotropy is not a defect but a division of labor that post-training can exploit.

1 INTRODUCTION

Supervised fine-tuning (SFT) and reinforcement learning (RL) are central to post-training large language models for reasoning (Ouyang et al., 2022; Shao et al., 2024). SFT learns from demonstrations, while reward-driven RL increases the probability of high-reward completions. Both operate on models with substantial linguistic and reasoning structure, making the organization of their existing representations a natural starting point for efficient adaptation. A characteristic feature of these representations is *anisotropy*: token representations exhibit preferential concentration along particular directions (Gao et al., 2019; Godey et al., 2024). Figure 1(a) illustrates this geometry. Understanding its role during post-training offers a representation-level perspective on how reasoning capabilities can be adapted.

Prior studies document anisotropy and frequency-associated structure in language-model representations (Martinez et al., 2024; Godey et al., 2024; Gao et al., 2019), while circuit-level analyses identify attention mechanisms and neurons that contribute to reasoning (Arora et al., 2026; Kim et al., 2025). Together, these findings establish representation structure as a useful lens for studying model behavior and motivate a closer examination of residual channels during post-training. We focus on two questions: *how do selected anisotropic channels and their complement evolve under SFT and RL, and how effectively can channel-wise reweighting adapt reasoning within a frozen backbone?* These questions connect the geometry of learned representations to the design of compact adaptation spaces.

In this study, we first characterize the occurrence and post-training dynamics of anisotropic residual channels. A label-free outlier rule on absolute mean post-block activations identifies a selected group that we use as a coherence proxy. In our Qwen2.5-1.5B diagnostics, this group comprises approximately 5% of post-SFT coordinates across the blocks displayed in Figure 1(b). Comparing independent training branches from the same pretrained checkpoint, we observe larger selected-

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.

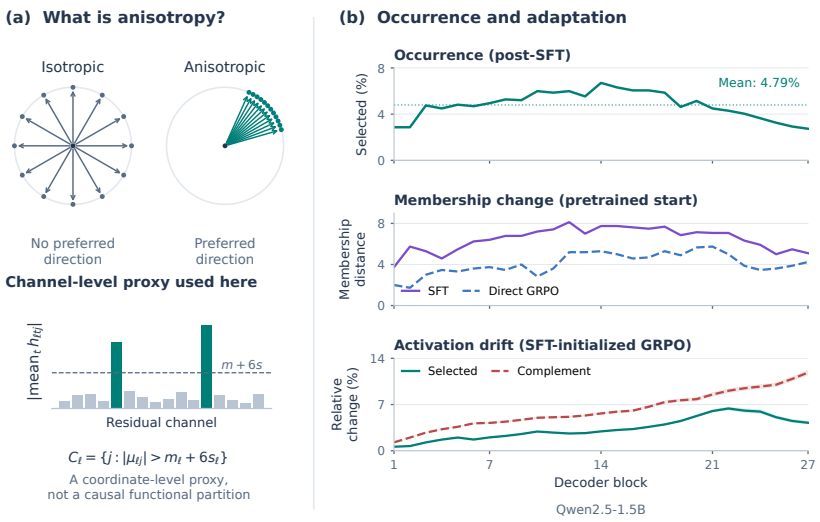

Figure 1: **Anisotropy and changes during post-training. (a)** Illustration of representations spread evenly across directions (isotropic) or concentrated around a preferred direction (anisotropic). The channel sketch highlights components whose average activations are unusually far from zero. **(b)** *Top:* these selected channels form a small group across the displayed layers. *Middle:* starting from the same pretrained model, supervised fine-tuning (SFT) changes which channels belong to this group more than direct reward-based training with GRPO in the compared runs. *Bottom:* in a separate GRPO run starting from SFT, the remaining channels change more relative to their starting activation levels. Shading shows uncertainty estimated by repeatedly resampling evaluation problems.

channel membership changes after SFT than after direct GRPO. A separate SFT-initialized full-model GRPO run exhibits larger baseline-relative activation changes in the complementary channels. Our rollout analysis also reports higher normalized subspace overlap between correct and incorrect rollouts in the selected group than in its complement (Figure 2). These observations provide a channel-level account of heterogeneous adaptation dynamics and motivate learning how to reweight the existing residual structure.

Building on this characterization, we introduce **SphereGate** to manipulate the relative weighting of residual channels during reasoning-oriented RL. Following the representation-level perspective of activation-scaling methods such as IA^3 (Liu et al., 2022), SphereGate learns one positive gain per channel on the full residual state after each decoder block, with all backbone weights held fixed. The gains are identity-initialized and have bounded log-values, yielding a compact diagonal adaptation operator. All channel gains are trainable in the default configuration, while the diagnostic partition supports analysis and restricted-gate ablations. Across four backbones on MATH-500, SphereGate exceeds the strongest TinyLoRA, LoRA-XS, or subnetwork baseline by **2.0–7.3 percentage points**. Complement-only adaptation retains nearly the all-channel gate's accuracy, further supporting the utility of channel-wise residual reweighting for reasoning.

Our contributions are threefold:

- **Characterizing anisotropic-channel dynamics.** We define a label-free mean-activation outlier partition and analyze its occurrence, rollout-subspace overlap, membership changes, and baseline-relative activation drift under SFT and GRPO.
- **Manipulating residual representations with a frozen backbone.** We introduce SphereGate, an identity-initialized positive full-residual gating method, and characterize its stability through an energy-weighted analysis under stated conditions.
- **Evaluating adaptation across scales and architectures.** Experiments on Qwen2.5-Instruct at 0.5B–7B and Llama-3-8B-Instruct demonstrate mathematical reasoning gains over the evaluated PEFT baselines, complemented by channel-restricted ablations.

2 Related Work

Parameter-efficient reinforcement learning. Recent work has shown that reinforcement learning for reasoning can be performed effectively by updating only a small subset of model parameters. TinyLoRA (Morris et al., 2026), LoRA-XS (Bałazy et al., 2024), and the subnetwork view of RL (Mukherjee et al., 2025) demonstrate that substantial reasoning gains can be obtained with highly parameter-efficient updates. These approaches, however, primarily identify the trainable parameter subspace using algebraic or parameter-space criteria, without explicitly considering the representational geometry of the model or the functional roles of individual representation channels (Dai et al., 2025; Liu et al., 2024). Classical parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2021) operate in weight space. Our experiments use Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a reinforcement-learning objective designed for reasoning with verifiable rewards. In contrast to weight-space approaches, SphereGate exploits the representation-space structure identified in our analysis and performs channel-wise gating of hidden representations, directing the adaptation budget toward reasoning-relevant channels while minimizing disruption to coherence-related channels.

Mechanistic interpretability and representation editing. A growing body of work has sought to identify the internal mechanisms responsible for reasoning in language models (Olsson et al., 2022; Du et al., 2026; Tang et al., 2026; Zhao et al., 2026). Recent studies analyze attention-based circuits and individual neurons to identify components that contribute causally to reasoning behavior (Arora et al., 2026; Kim et al., 2025), providing evidence that reasoning is mediated by structured and often sparse subsets of model components. A parallel line of work studies how model behavior can be controlled through interventions on internal representations (Zou et al., 2023), including steering via activation differences (Turner et al., 2024; Panickssery et al., 2023), learned representation transformations such as ReFT and LoReFT (Wu et al.). These analyses, however, primarily operate at the level of attention heads, neurons, or architectural components rather than the residual channel structure. Our work instead investigates the residual representation at the channel level, identifying a distinction between reasoning-oriented channels and a separate set of highly anisotropic channels associated with coherent behavior, and directly exploits this structure during RL.

3 Analysis

We examine how correct and incorrect trajectories align and how SFT and GRPO modify their representations. We use anisotropic channels as a coherence proxy and refer to their complement as reasoning channels. For a matrix A, the Frobenius norm is

$$\|A\|_F = \left(\sum_{i,j} |A_{ij}|^2 \right)^{1/2}. \tag{1}$$

3.1 Diagonal-Gate versus Full-GRPO Similarity.

In a separate matched-budget experiment, we compare non-anisotropic diagonal-gate GRPO with full-model GRPO, both trained independently from the same pretrained Qwen2.5-1.5B checkpoint without SFT initialization. Each run uses 1,000 rollout batches and 64,000 trajectories. Figure 2(d) reports the mean per-token cosine similarity of their post-block residual states on identical teacher-forced prefixes from 128 held-out gold solutions, pooled over 27,742 prediction positions at each decoder block. States are compared in their native coordinates without fitted alignment; the final block is measured before the final RMSNorm. Similarity is high in early blocks and decreases slightly motivating the existence of a diagonal only type of gate for reinforcement learning algorithms as post training for Large Language models.

3.2 Correct and Wrong Subspace Alignment

We compare the representation subspaces of correct and incorrect rollouts, using normalized Frobenius overlap within each channel group. We generate correct and incorrect rollouts for each prompt and then independently aggregate them to construct the correct and incorrect subspaces with spectral

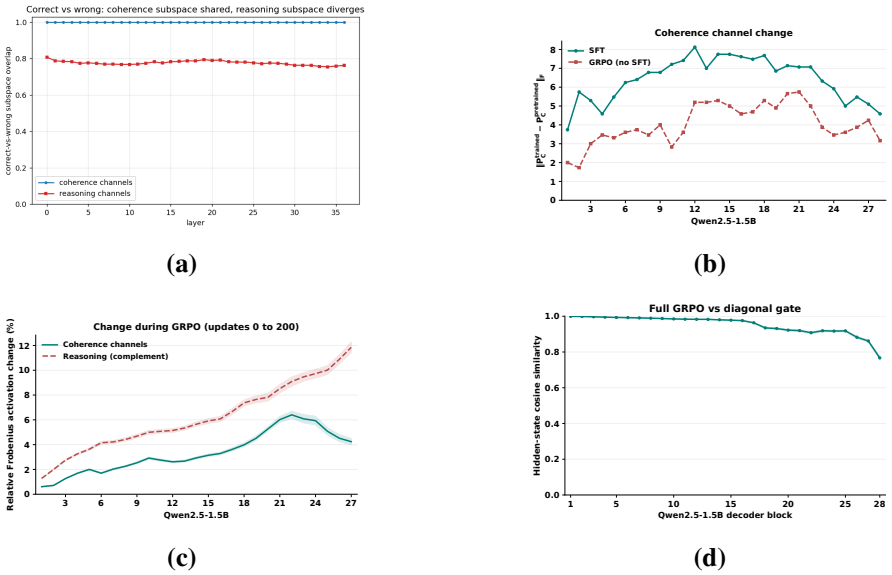

Figure 2: **Coherence and reasoning representations.** **(a)** Correct and incorrect rollouts have higher subspace overlap in coherence channels than in their complement. **(b)** SFT produces larger coherence-channel membership changes than direct GRPO relative to the same pretrained model. **(c)** During SFT-initialized GRPO, complementary reasoning channels undergo larger baseline-relative activation changes than coherence channels. **(d)** Hidden-state cosine similarity between non-anisotropic diagonal-gate GRPO and full-model GRPO, both initialized from the same pretrained checkpoint, across decoder blocks 1–28. Shading denotes pointwise 95% problem-bootstrap confidence intervals.

decomposition, respectively. For group $g \in \{C, R\}$, let U_g^+ and U_g^- be orthonormal bases for the correct and incorrect subspaces, with nonzero ranks r_g^+ and r_g^-. Their overlap is

$$\mathcal{O}_g = \frac{\left\|(U_g^+)^\top U_g^-\right\|_F}{\sqrt{\min(r_g^+, r_g^-)}}. \tag{2}$$

Figure 2(a) shows nearly complete alignment in the coherence channels, whereas the complementary reasoning channels exhibit lower overlap. The paired trajectories thus show stronger subspace alignment in the coherence group despite their different outcomes.

3.3 Coherence-Channel Change under SFT and GRPO

We compare independent SFT and full-model GRPO branches initialized from the same pretrained model, with no SFT initialization for the GRPO branch. Figure 2(b) reports the Frobenius distance between each trained model's coherence-channel projector and the pretrained projector, estimated on shared calibration prompts. Writing $C^{(t)}$ for the coherence-channel set and $P_C^{(t)} = \mathrm{diag}(\mathbf{1}_{C^{(t)}})$ for its projector, we measure

$$\begin{aligned} D_C^{(t)} &= \left\| P_C^{(t)} - P_C^{(\mathrm{pre})} \right\|_F \\ &= \sqrt{\left| C^{(t)} \triangle C^{(\mathrm{pre})} \right|}. \end{aligned} \tag{3}$$

Here $t \in \{\mathrm{SFT}, \mathrm{GRPO}\}$, pre denotes pretraining, and $\triangle$ is symmetric set difference. SFT shows larger coherence-channel membership changes than direct GRPO for these checkpoints. Both branches modify the identified channels, but the direct-GRPO changes are smaller.

3.4 Coherence versus Reasoning Changes during GRPO

We next examine a separate full-model GRPO run initialized from SFT, keeping the initial channel partition fixed. Figure 2(c) shows the Frobenius norm of activation change, normalized by each group's initial activation norm, using identical completion tokens before and after GRPO. Let $H^{(0)}$ and $H^{(1)}$ be the initial and final activations, and $P_g^{(0)}$ the fixed initial group projector, with $P_R^{(0)} = I - P_C^{(0)}$. Raw and percentage-relative changes are

$$F_g = \left\| (H^{(1)} - H^{(0)}) P_g^{(0)} \right\|_F,$$
$$A_g = 100 \, \frac{F_g}{\left\| H^{(0)} P_g^{(0)} \right\|_F}, \qquad g \in \{C, R\}. \tag{4}$$

The plot reports A_g. The complementary reasoning channels show larger relative changes, while coherence-channel changes remain smaller but nonzero. The shaded bands represent pointwise 95% problem-bootstrap confidence intervals.

4 Methodology

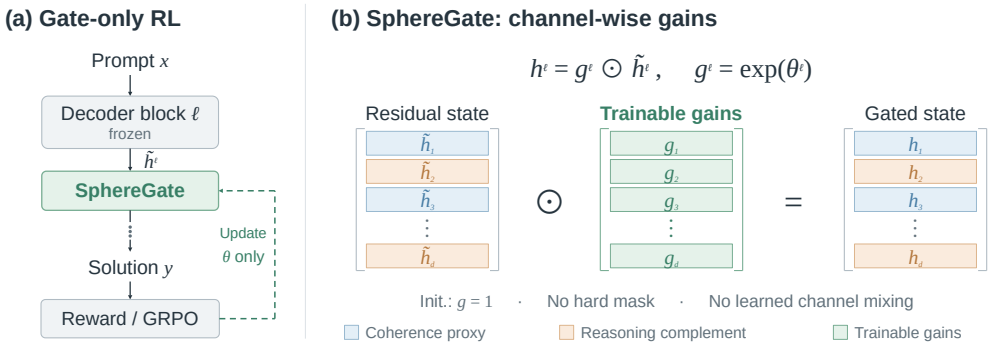

Figure 3: The General Framework for Our Proposed Method, SphereGate.

4.1 SphereGate

Consider a decoder block l with residual input $h^{l-1} \in \mathbb{R}^d$. A standard Transformer block writes into the residual stream through an attention sublayer followed by an MLP sublayer:

$$a^l = \mathrm{Attn}^l(\mathrm{LN}(h^{l-1})), \tag{5}$$
$$u^l = h^{l-1} + a^l, \tag{6}$$
$$m^l = \mathrm{MLP}^l(\mathrm{LN}(u^l)), \tag{7}$$
$$\tilde{h}^l = u^l + m^l, \tag{8}$$

where $\tilde{h}^l$ is the full residual state after the MLP write, and LN denotes layer normalization (modern models typically use RMSNorm). SphereGate then applies a learned per-channel gain directly to this residual state:

$$h^l = g^l \odot \tilde{h}^l, \qquad g^l = \exp(\theta^l). \tag{9}$$

During training, only θ^l is trainable; all backbone parameters remain frozen.

4.2 Gradient Analysis

Let L denote the number of decoder blocks and d_ℓ the residual dimension at block ℓ. For token position t, SphereGate maps the pre-gate state $\widetilde{h}_{\ell t} \in \mathbb{R}^{d_\ell}$ to

$$h_{\ell t} = \exp(\theta_\ell) \odot \widetilde{h}_{\ell t},$$

where θ_ℓ contains the channel-wise log-gains. Within their bounds $|\theta_{\ell k}| < c$, with $c > 0$, the gradient of a differentiable loss $\mathcal{L}$ is

$$\nabla_{\theta_\ell} \mathcal{L} = \sum_{t \in \mathcal{T}} \frac{\partial \mathcal{L}}{\partial h_{\ell t}} \odot h_{\ell t}, \tag{10}$$

where $\mathcal{T}$ contains the nonpadding token positions in the fixed training batch. Thus, gain gradients aggregate activation-weighted loss sensitivities.

Channel partition. As mentioned in the figure 1, we identify anisotropic channels using a fixed calibration pass. Let $h^{\text{cal}}_{\ell t k}$ denote the post-block activation of channel k at calibration position t, measured before final model normalization at the last block. For the nonpadding positions $\mathcal{T}_{\text{cal}}$, with $N_{\text{cal}} = |\mathcal{T}_{\text{cal}}|$, define

$$\mu_{\ell k} = \frac{1}{N_{\text{cal}}} \sum_{t \in \mathcal{T}_{\text{cal}}} h^{\text{cal}}_{\ell t k}, \qquad v_{\ell k} = |\mu_{\ell k}|.$$

Using the median and unscaled median absolute deviation,

$$m_\ell = \text{median}_k \, v_{\ell k}, \qquad s_\ell = \text{median}_k \, |v_{\ell k} - m_\ell|,$$

we select

$$C_\ell = \{k : v_{\ell k} > m_\ell + 6 s_\ell\}, \qquad R_\ell = \{1, \ldots, d_\ell\} \setminus C_\ell.$$

Index all layer-channel pairs by $j \in \{1, \ldots, p\}$, where $p = \sum_{\ell=1}^{L} d_\ell$. The sets C and R collect the selected and complementary coordinates across layers and remain fixed during the analysis. Default SphereGate training optimizes gains on all coordinates.

Energy-weighted stability. Let π_θ denote the gate-adapted policy. For a fixed rollout batch, consider the maximization objective

$$J(\theta) = U(\theta) - \beta K(\theta), \qquad \beta > 0,$$

where U is the length-normalized GRPO reward surrogate and K is the KL penalty to a frozen reference policy π_{ref}. Samples, advantages, and sequence lengths are fixed during this optimization stage.

Let θ_0 be the rollout-policy log-gains, θ_1 a subsequent iterate, and $\Delta = \theta_1 - \theta_0$. Assume that the connecting segment lies inside the gain bounds and that J is twice continuously differentiable on a neighborhood of the segment. Define the baseline channel energies and their diagonal matrix by

$$E_j = \sum_{t \in \mathcal{T}} h_{tj}(\theta_0)^2 > 0, \qquad D = \text{diag}(E_1, \ldots, E_p),$$

where h_{tj} denotes the activation at the layer-channel pair indexed by j.

Write $z = D^{1/2}\Delta$ for the energy-weighted log-gain displacement. Define the selected group's initial gradient magnitude and final gradient residual as

$$\begin{aligned} \varepsilon_C &= \|D_C^{-1/2} \nabla_C J(\theta_0)\|_2, \\ \tau_C &= \|D_C^{-1/2} \nabla_C J(\theta_1)\|_2. \end{aligned} \tag{11}$$

Subscripts denote coordinate restrictions or matrix blocks; in particular, D_C is the principal block indexed by C. The norm $\|\cdot\|_2$ is Euclidean for vectors and spectral for matrices.

Define the integrated normalized curvature

$$\mathcal{H} = D^{-1/2} \left[-\int_0^1 \nabla^2 J(\theta_0 + s\Delta) \, ds \right] D^{-1/2}. \tag{12}$$

Proposition 1 (Energy-weighted gate stability)**.** *Suppose that, for $\lambda > 0$ and $\rho \geq 0$,*

$$\mathcal{H}_{CC} \succeq \lambda I_{|C|}, \qquad \|\mathcal{H}_{CR}\|_2 \leq \lambda \rho,$$

where $\succeq$ denotes positive-semidefinite ordering and $I_{|C|}$ is the identity on the selected coordinates. Then

$$\|z_C\|_2 \leq B_C := \frac{\varepsilon_C + \tau_C}{\lambda} + \rho \|z_R\|_2, \tag{13}$$

$$|\Delta_j| \leq \frac{B_C}{\sqrt{E_j}}, \qquad j \in C. \tag{14}$$

Table 1: Performance on MATH-500 under GRPO using greedy decoding. All values are absolute accuracy scores on a 0–100 scale. Standard deviations are estimated across three training seed checkpoints.

Method	Qwen2.5-0.5B-It	Qwen2.5-3B-It	Qwen2.5-7B-It	Llama-3-8B-It
Base Model	32.5 ±0.82	64.5 ±0.71	70.0 ±0.80	33.5 ±0.76
Full GRPO	41.6 ±0.72	69.2 ±0.60	77.2 ±0.61	37.4 ±0.65
IA^3	31.7 ±1.02	63.0 ±0.88	68.6 ±0.94	32.4 ±0.93
TinyLoRA	33.5 ±0.74	68.0 ±0.63	75.0 ±0.58	35.0 ±0.72
LoRA-XS	35.0 ±0.68	67.0 ±0.57	77.0 ±0.52	36.0 ±0.66
Subnetwork	34.5 ±0.91	67.0 ±0.76	73.0 ±0.70	34.0 ±0.84
SphereGate	**42.3 ±0.61**	**70.0 ±0.54**	**79.5 ±0.46**	**38.0 ±0.58**

The bound separates selected-group optimization pressure from cross-group coupling. For a fixed B_C, it provides an inverse-square-root energy envelope on coordinate-wise log-gain movement. Moreover, if $\|z_R\|_2 > 0$ and

$$\frac{\varepsilon_C + \tau_C}{\lambda} \leq \alpha \|z_R\|_2, \qquad \alpha \geq 0, \quad \alpha + \rho < 1,$$

then

$$\|z_C\|_2 \leq (\alpha + \rho)\|z_R\|_2 < \|z_R\|_2.$$

This gives a precise criterion for adaptation to concentrate in the complementary group under the energy-weighted metric. Appendix A.3 provides the proof, the corresponding direct representation-displacement bound, and explicit connections to the GRPO reward signal and KL curvature.

5 Results

5.1 Experimental Setup

Models and data. We evaluate SphereGate on Qwen2.5-Instruct (0.5B, 1.5B, 3B, and 7B), Llama-3-8B-Instruct Grattafiori et al. (2025), and Qwen2.5-Math-7B Yang et al. (2024). Training data comprise MATH-train (Hendrycks et al., 2021), DAPO-Math-17k Yu et al. (2026), and MIX10k (an equal mixture of 5k MATH and 5k DAPO).

GRPO. We use a PPO-clip surrogate with clip-eps 0.2, $\mu = 2$ inner epochs, and group-normalized advantages (std_adv=True, len_norm=True). The KL coefficient is 0.01; the gate-disabled model serves as the identity reference, avoiding a second model copy. Training uses 90 steps, 64 prompts and 8 completions per prompt (512 rollouts per step), temperature 1.0, top-p 1.0, and a 1024-token cap (4096 for the Math model).

Baselines. We compare with TinyLoRA (Morris et al., 2026), LoRA-XS (Bałazy et al., 2024), random subnetwork adaptation (Mukherjee et al., 2025), IA^3, LoRA (Hu et al., 2021), and full fine-tuning. SphereGate, LoRA-XS, TinyLoRA, and Subnetwork each train ≈0.100M parameters, with learning rates of 1e-3, 1e-3, 2e-4, and 1e-4, respectively. SphereGate trains all channels with θ initialized to 0 and clamped at $c = 0.7$; LoRA-XS and TinyLoRA use SVD-rank 64, and Subnetwork uses subnet-k 10000. IA^3 uses a learning rate of $1e-5$ and learned scaling vectors on key, value, and feed-forward projections; full fine-tuning uses 1e-6. All methods share prompts, rewards, held-out splits, and rollout budgets.

5.2 MATH-500 Performance

SphereGate achieves the highest reported greedy MATH-500 accuracy on every backbone (Table 1). On Qwen2.5-0.5B, its 42.3 exceeds the strongest evaluated PEFT baseline, LoRA-XS (35.0), by 7.3 points; the corresponding 7B margin is 2.5 points. Accuracy improves over the base models from 64.5 to 70.0 on Qwen2.5-3B, 70.0 to 79.5 on Qwen2.5-7B, and 33.5 to 38.0 on Llama-3-8B-Instruct.

Table 2: Mathematical reasoning performance on AMC23 and Minerva Math using Avg@16. Standard deviations are estimated across three training seed checkpoints.

Method	AMC23		Minerva Math	
	Qwen2.5-7B-It	Llama-3-8B-It	Qwen2.5-7B-It	Llama-3-8B-It
Base Model	51.9 $\pm$1.38	8.8 $\pm$0.47	33.66 $\pm$0.72	12.39 $\pm$0.61
Full GRPO	53.6 $\pm$1.00	10.2 $\pm$0.34	36.1 $\pm$0.70	14.1 $\pm$0.70
IA^3	50.8 $\pm$1.47	8.4 $\pm$0.52	32.90 $\pm$0.81	11.80 $\pm$0.78
LoRA-XS	52.8 $\pm$1.11	9.2 $\pm$0.39	34.85 $\pm$0.61	12.34 $\pm$0.74
TinyLoRA	52.2 $\pm$1.27	9.4 $\pm$0.22	34.73 $\pm$0.58	11.08 $\pm$0.86
Subnetwork	53.7 $\pm$1.13	9.6 $\pm$0.28	34.87 $\pm$0.86	12.02 $\pm$0.88
SphereGate	**55.0** $\pm$**0.82**	**10.0** $\pm$**0.31**	**36.28** $\pm$**0.66**	**13.88** $\pm$**0.68**

5.3 Mathematical Reasoning Performance

Table 2 reports mean accuracy over 16 sampled responses (Avg@16) on AMC23 Cao et al. (2025) and Minerva Math Lewkowycz et al. (2022) for Qwen2.5-7B-Instruct and Llama-3-8B-Instruct. SphereGate leads the evaluated PEFT baselines in all four settings. On Qwen2.5-7B AMC23, it scores 55.0%, versus 53.7% for Subnetwork and 51.9% for the base model (a 3.1-point gain). On Llama-3-8B Minerva Math, it gains 1.49 points over the base model, compared with flat or lower scores for TinyLoRA and LoRA-XS.

5.4 Competition and Olympiad Benchmarks

On Qwen2.5-7B-Math-Instruct, SphereGate reaches 23.33% on AIME24 and 44.0% on OlympiadBench using maj@16 (Table 3), supporting out-of-distribution transfer. On AIME24, it improves the base score (16.67%) by 6.66 points and exceeds LoRA-XS (20.00%), while TinyLoRA and Subnetwork match the base score.

Table 3: Transfer performance on AIME24 and OlympiadBench for Qwen2.5-7B-Math-Instruct. AIME24 and OlympiadBench is evaluated using maj@16, All values are percentages.

Method	AIME24	OlympiadBench
Base Model	16.67	41.5
Full GRPO	26.67	44.5
IA^3	16.67	40.2
LoRA-XS	20.00	43.0
TinyLoRA	16.67	43.0
Subnetwork	16.67	42.0
SphereGate	**23.33**	**44.0**

5.5 Channel Ablations: Reasoning Adaptation and General-Text Prediction

We distinguish restricted-gain training from inference-time activation removal (Table 4). On Qwen2.5-3B-Instruct, excluded gains remain fixed at one, preserving their activations. Complement-only GRPO reaches 69.4 on MATH-500, within 0.6 percentage points of all-channel SphereGate (70.0).

For general-text prediction, we remove count-matched channel groups from pretrained Qwen2.5-1.5B and its MATH-SFT checkpoint. Masks estimated from 256 MATH training prompts select 4.19% and 4.36% of coordinates on average, respectively. Each block's selected group is compared with an equal number of uniformly sampled complementary channels, averaging control perplexities across five random-mask seeds. All conditions score the same 299,077 WikiText-2 test targets with context length 1,024 and stride 512, without retraining. Per-token residual-norm restoration to the value immediately before masking additionally controls scale reduction. Anisotropic-channel removal causes substantially greater prediction damage even with this control: SFT perplexity rises from 10.44 to 2.01×10^6, versus 35.06 for matched complementary removal.

(a) Restricted-gain training: Qwen2.5-3B-Instruct	
Gate variant	MATH-500 greedy (%)
Baseline (un trained)	64.5
SphereGate (all channels)	70.0
Anisotropic-only gate	65.1
Random Channel gate	66.9
Matched Complement-only gate	67.5
Complement-only gate	69.4

(b) Activation removal: Qwen2.5-1.5B			
Removed channels	Scaling	Pretrained PPL	SFT PPL
None	–	8.84	10.44
Anisotropic	None	3.91×10^6	2.44×10^6
Matched complement	None	31.18	37.96
Anisotropic	Norm-restored	2.73×10^6	2.01×10^6
Matched complement	Norm-restored	29.21	35.06

Table 4: **Complementary channel ablations.** (a) MATH-500 accuracy after GRPO with different trainable channel groups; higher is better. Excluded gains remain at one, so their activations are not removed. (b) WikiText-2 perplexity under inference-time activation removal; lower is better. Anisotropic and matched-complement conditions remove the same number of coordinates in every decoder block. Matched-complement perplexities are averaged across five random-mask seeds. Full-test perplexities use BF16 inference and FP64 loss accumulation. Norm restoration is an ablation control, not part of the SphereGate training operator.

These findings are consistent with selected channels being important for general-text prediction and complementary gains enabling effective reasoning adaptation, supporting anisotropy as an operational coherence proxy. They do not establish exclusive functional localization: model sizes differ, restricted gates have unequal parameter counts, and count matching with norm restoration equalizes neither removed activation energy nor directional displacement. Perplexity measures next-token prediction rather than semantic coherence directly.

6 Conclusion

SphereGate demonstrates that learning diagonal channel-wise gains is sufficient for effective GRPO adaptation of a frozen backbone, achieving strong mathematical reasoning performance across Qwen2.5 and Llama-3 models. Our analyses reveal a channel-level separation between coherence and reasoning: anisotropic residual channels predominantly support coherence, while complementary channels provide the primary capacity for reasoning adaptation.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT, Claude, and Claude Code to implement methods, write and edit code, and assist in drafting the manuscript. We did not use generative AI tools to generate synthetic data, formulate , or propose our core analysis, methodology, or experimental design; these were carried out by the authors. We have reviewed all AI-assisted work: code was checked against expected behavior and the underlying data, and all AI-drafted text and result interpretations were verified against experimental outputs and revised where necessary. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Aryaman Arora, Zhengxuan Wu, Jacob Steinhardt, and Sarah Schwettmann. Language model circuits are sparse in the neuron basis. *arXiv preprint arXiv:2601.22594*, 2026.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. Lora-xs: Low-rank adaptation with extremely small number of parameters. *arXiv preprint arXiv:2405.17604*, 2024.

Lang Cao, Yingtian Zou, Chao Peng, Renhong Chen, Wu Ning, and Yitong Li. Step guided reasoning: Improving mathematical reasoning using guidance generation and step reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21112–21129, 2025.

Gaole Dai, Chun-Kai Fan, Yiming Tang, Zhi Zhang, Yuan Zhang, Yulu Gan, Qizhe Zhang, Cheng-Ching Tseng, Shanghang Zhang, and Tiejun Huang. SAN: Hypothesizing long-term synaptic development and neural engram mechanism in scalable model's parameter-efficient fine-tuning. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=CeTXRRAdFi`.

Lingheng Du, Yiming Tang, Xufeng Duan, and Dianbo Liu. What does an llm learn from reinforcement learning? a mechanistic interpretability perspective with fixed-sae track. *arXiv preprint arXiv:2609.15064*, 2026.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35–48, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. 2025.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Geonhee Kim, Marco Valentino, and Andre Freitas. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10074–10095, 2025.

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.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in neural information processing systems*, 35:1950–1965, 2022.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

Richard Diehl Martinez, Zébulon Goriely, Andrew Caines, Paula Buttery, and Lisa Beinborn. Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5999–6011, 2024.

John X Morris, Niloofar Mireshghallah, Mark Ibrahim, and Saeed Mahloujifar. Learning to reason in 13 parameters. *arXiv preprint arXiv:2602.04118*, 2026.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models, 2025. *URL https://arxiv. org/abs/2505.11711*, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. *URL https://arxiv. org/abs/2312.06681*, 3, 2023.

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.

Yiming Tang, Harshvardhan Saini, Samyak Jha, Huaming Chen, Xufeng Duan, and Dianbo Liu. Capsule lens: Locating and tracking concept geometry in model representations. *arXiv preprint arXiv:2609.05575*, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. *URL https://arxiv. org/abs/2308.10248*, 2308, 2024.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models, 2024. *URL https://arxiv. org/abs/2404.03592*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Haiyan Zhao, Zirui He, Yiming Tang, Fan Yang, Ali Payani, Dianbo Liu, and Mengnan Du. Rep2text: Decoding full text from a single llm token representation, 2026. *URL https://arxiv. org/abs/2511.06571*, 2511, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Appendix

This appendix provides the formal derivations and geometric proofs supporting the empirical analysis in Section 3 and the parameter-efficient adaptation dynamics of SphereGate in Section 4.

A.1 Forward-KL Demonstration Coverage vs. Advantage-Weighted Optimization

Let $x \sim d$ be a prompt sampled from the task distribution, $q(y \mid x)$ be the demonstration distribution, and $\pi_\theta(y \mid x)$ be an autoregressive policy over completions $y \in \mathcal{Y}$. We assume a finite completion space, a fixed reference policy π_{ref} with full support, and a bounded reward function $R(x, y)$.

Proposition 2 (SFT Minimizes Expected Forward-KL)**.** *The sequence negative log-likelihood loss satisfies*

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{x\sim d,\, y\sim q}\big[\log \pi_\theta(y \mid x)\big] = H_q(Y \mid X) + \mathbb{E}_{x\sim d}\big[D_{\text{KL}}\big(q(\cdot \mid x) \,\|\, \pi_\theta(\cdot \mid x)\big)\big]. \quad (15)$$

Because the conditional data entropy $H_q(Y \mid X)$ is invariant to θ, maximum likelihood estimation under teacher forcing is strictly equivalent to minimizing the expected forward Kullback–Leibler divergence.

Proof. By identity, $-\log \pi_\theta(y \mid x) = -\log q(y \mid x) + \log \frac{q(y|x)}{\pi_\theta(y|x)}$. Taking expectations with respect to $x \sim d$ and $y \sim q(\cdot \mid x)$ yields Equation equation 15. Since $\log \pi_\theta(y \mid x) = \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t})$, the token-level cross-entropy loss sums directly to this sequence divergence. □

Proposition 3 (Forward-KL Strongly Penalizes Dropping Empirical Modes)**.** *Fix a prompt x and let $S \subset \mathcal{Y}$ be any completion subset. Let $a = q(S \mid x) \in (0, 1)$ and $b = \pi_\theta(S \mid x) \in (0, 1)$. Then:*

$$D_{\text{KL}}\big(q(\cdot \mid x) \,\|\, \pi_\theta(\cdot \mid x)\big) \geq a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b}. \quad (16)$$

Consequently, for any data-supported region ($a > 0$), $D_{\text{KL}}(q \,\|\, \pi_\theta) \to +\infty$ as $b \to 0$.

Proof. Suppressing conditioning on x, decompose the divergence across S and S^c:

$$\begin{aligned} D_{\text{KL}}(q \,\|\, \pi_\theta) = a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b} \\ + aD_{\text{KL}}\big(q(\cdot \mid S) \,\|\, \pi_\theta(\cdot \mid S)\big) + (1 - a)D_{\text{KL}}\big(q(\cdot \mid S^c) \,\|\, \pi_\theta(\cdot \mid S^c)\big). \end{aligned} \quad (17)$$

Nonnegativity of the conditional KL divergences gives the lower bound in Equation equation 16. The limit follows immediately since $\lim_{b\to 0^+} a \log(a/b) = +\infty$. □

Remark (Scope of Forward-KL). Proposition 3 proves that SFT places an infinite penalty on omitting demonstration modes. This explains why SFT fits demonstration frequencies rather than explicitly distinguishing correctness, establishing the shared coherence substrate.

Proposition 4 (KL-Regularized Policy Optimization as Reverse-KL Projection)**.** *Consider the sequence-level KL-regularized reward objective:*

$$J_\beta(\pi) = \mathbb{E}_{x\sim d}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}[R(x, y)] - \beta D_{\text{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)\big)\right], \quad \beta > 0. \quad (18)$$

Define the partition function $Z_\beta(x) = \sum_y \pi_{\text{ref}}(y \mid x) \exp(R(x, y)/\beta)$ and the tilted target distribution $\pi^*_\beta(y \mid x) = \frac{1}{Z_\beta(x)} \pi_{\text{ref}}(y \mid x) \exp(R(x, y)/\beta)$. Then:

$$J_\beta(\pi) = \beta \mathbb{E}_{x\sim d}\big[\log Z_\beta(x)\big] - \beta \mathbb{E}_{x\sim d}\big[D_{\text{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi^*_\beta(\cdot \mid x)\big)\big]. \quad (19)$$

Proof. Expanding the divergence between π and π^*_β for a fixed prompt x:

$$\begin{aligned} D_{\text{KL}}(\pi \,\|\, \pi^*_\beta) &= \mathbb{E}_{y\sim\pi}\left[\log \frac{\pi(y \mid x)}{\pi_{\text{ref}}(y \mid x)} - \frac{R(x, y)}{\beta} + \log Z_\beta(x)\right] \\ &= D_{\text{KL}}(\pi \,\|\, \pi_{\text{ref}}) - \frac{1}{\beta}\mathbb{E}_{y\sim\pi}[R(x, y)] + \log Z_\beta(x). \end{aligned} \quad (20)$$

Multiplying through by $-\beta$ and taking expectations over $x \sim d$ yields Equation equation 19. □

Surrogate Gradient of GRPO. In practical Group Relative Policy Optimization (Shao et al., 2024), $G \geq 2$ trajectories $\{y_i\}_{i=1}^{G}$ are sampled from the current rollout checkpoint $\mu = \pi_{\theta_{\text{old}}}$. Evaluating the token-averaged PPO-clip surrogate at $\theta = \theta_{\text{old}}$, the policy ratio equals 1 and clipping is inactive, yielding the local gradient:

$$\nabla_\theta \widehat{J}_{\text{rew}}(\theta)\Big|_{\theta_{\text{old}}} = \frac{1}{G}\sum_{i=1}^{G} \frac{\widehat{A}_i}{|y_i|} \nabla_\theta \log \pi_\theta(y_i \mid x)\Bigg|_{\theta_{\text{old}}}, \tag{21}$$

where $\widehat{A}_i = \frac{R_i - \bar{R}}{s_R + \delta}$ is the group-standardized advantage (s_R being the sample standard deviation of rewards).

A.2 Reward-Selective Gradients under Group Advantage Centering

Let $s_i = \frac{1}{|y_i|} \nabla_\theta \log \pi_\theta(y_i \mid x)\Big|_{\theta_{\text{old}}}$ be the length-normalized score vector of completion i, with group mean $\bar{s} = \frac{1}{G}\sum_{i=1}^{G} s_i$. The unregularized reward gradient is $g_{\text{rew}} = \frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i s_i$.

Proposition 5 (Centered Update as Reward–Score Covariance)**.** *The local reward gradient under group normalization is proportional to the sample covariance between scalar rewards and score vectors:*

$$g_{\text{rew}} = \frac{1}{G(s_R + \delta)} \sum_{i=1}^{G} (R_i - \bar{R})(s_i - \bar{s}). \tag{22}$$

Proof. Substituting $\widehat{A}_i = (R_i - \bar{R})/(s_R + \delta)$:

$$g_{\text{rew}} = \frac{1}{G(s_R + \delta)} \sum_{i=1}^{G} (R_i - \bar{R}) s_i. \tag{23}$$

Because the centered rewards sum to zero ($\sum_{i=1}^{G}(R_i - \bar{R}) = 0$), we have $\sum_{i=1}^{G}(R_i - \bar{R})\bar{s} = 0$. Subtracting this term yields Equation equation 22. □

Corollary 5.1 (Shared Score Invariance)**.** *Decompose each completion score into a component shared across the rollout group and a completion-specific residual: $s_i = c + d_i$. For any parameter direction v, if $v^\top s_i$ is invariant across all completions in the group ($v^\top d_i = 0$), then:*

$$v^\top g_{\text{rew}} = 0. \tag{24}$$

Proof. Projecting Equation equation 22 onto v:

$$v^\top g_{\text{rew}} = \frac{1}{G(s_R + \delta)} \sum_{i=1}^{G} (R_i - \bar{R})\big(v^\top s_i - v^\top \bar{s}\big). \tag{25}$$

If $v^\top s_i = v^\top c$ for all $i \in \{1, \ldots, G\}$, then $v^\top \bar{s} = v^\top c$. Every term in the sum vanishes identically. □

Remark (Scope of the Covariance Form). Corollary 5.1 provides an objective-level motivation for why coherence representations remain stable during RL: features common to both successful and unsuccessful completions produce zero reward gradient under group centering. Note that shared activations do not strictly imply identical score vectors (which also depend on downstream sensitivities), but this establishes a strong structural bias against modifying universally shared sequence components.

A.3 Local Geometry and Selective Adaptation of Residual Gates

We derive the geometry of positive residual gates, prove Proposition 1, and express its stability conditions through the GRPO reward signal and the KL-regularized objective. Throughout, $\|\cdot\|_F$ denotes the Frobenius norm.

Positive residual gates. For a matrix of pre-gate token activations $\widetilde{H}_\ell$, the layer-ℓ gate gives

$$H_\ell = \widetilde{H}_\ell G_\ell, \qquad G_\ell = \operatorname{diag}\big(\exp(\theta_\ell)\big).$$

At initialization, $\theta_\ell = 0$ and $G_\ell = I$. For $|\theta_{\ell k}| \le c$, every diagonal entry lies in $[e^{-c}, e^{c}]$. Hence G_ℓ is invertible, and the gate preserves the entrywise signs, zero pattern, and rank of its pre-gate input. Furthermore,

$$\begin{aligned} e^{-c}\|\widetilde{H}_\ell\|_F \le \|H_\ell\|_F \le e^{c}\|\widetilde{H}_\ell\|_F, \\ \|H_\ell - \widetilde{H}_\ell\|_F \le (e^{c}-1)\|\widetilde{H}_\ell\|_F. \end{aligned} \tag{26}$$

These inequalities follow by applying the scalar gain bounds to each column.

Within the interior of the log-gain bounds,

$$\frac{\partial h_{\ell tk}}{\partial \theta_{\ell k}} = \widetilde{h}_{\ell tk} e^{\theta_{\ell k}} = h_{\ell tk}.$$

The chain rule, summed over token positions, therefore gives

$$\frac{\partial \mathcal{L}}{\partial \theta_{\ell k}} = \sum_{t \in \mathcal{T}} \frac{\partial \mathcal{L}}{\partial h_{\ell tk}} h_{\ell tk},$$

establishing Equation equation 10.

Direct gate geometry. Let $\Delta = \theta_1 - \theta_0$. Evaluate the direct gate intervention by fixing each pre-gate input at its value under θ_0. For the flattened layer-channel index j,

$$\Delta h_{tj}^{\mathrm{direct}} = h_{tj}(\theta_0)\big(e^{\Delta_j} - 1\big). \tag{27}$$

Consequently, for any coordinate group g,

$$\|\Delta H_g^{\mathrm{direct}}\|_F^2 = \sum_{j \in g} E_j \big(e^{\Delta_j} - 1\big)^2, \tag{28}$$

where $E_j = \sum_{t \in \mathcal{T}} h_{tj}(\theta_0)^2$. For groups spanning multiple layers, the squared norm sums over their layer-token-channel entries.

As $\Delta_g \to 0$, Taylor expansion yields

$$\|\Delta H_g^{\mathrm{direct}}\|_F^2 = \Delta_g^\top D_g \Delta_g + O\big(\|\Delta_g\|_2^3\big). \tag{29}$$

Thus, D is the local quadratic metric of direct gate-induced representation displacement.

A finite-displacement bound follows from the mean value theorem. Define $\delta_g = \|\Delta_g\|_\infty = \max_{j \in g} |\Delta_j|$. For every $|u| \le \delta_g$,

$$e^{-\delta_g}|u| \le |e^u - 1| \le e^{\delta_g}|u|.$$

Substitution into Equation equation 28 gives

$$e^{-\delta_g}\|D_g^{1/2}\Delta_g\|_2 \le \|\Delta H_g^{\mathrm{direct}}\|_F \le e^{\delta_g}\|D_g^{1/2}\Delta_g\|_2. \tag{30}$$

Proof of energy-weighted gate stability.

Proof of Proposition 1. Define the full normalized gradient vectors

$$a = D^{-1/2}\nabla J(\theta_0), \qquad e = D^{-1/2}\nabla J(\theta_1).$$

By Equation equation 11, $\|a_C\|_2 = \varepsilon_C$ and $\|e_C\|_2 = \tau_C$.

The fundamental theorem of calculus gives

$$\nabla J(\theta_1) - \nabla J(\theta_0) = \left[\int_0^1 \nabla^2 J(\theta_0 + s\Delta)\, ds\right] \Delta.$$

Multiplying by $D^{-1/2}$ and using $z = D^{1/2}\Delta$ yields

$$\mathcal{H}z = a - e. \tag{31}$$

Taking the selected-coordinate block,

$$\mathcal{H}_{CC}z_C = a_C - e_C - \mathcal{H}_{CR}z_R.$$

Since $\mathcal{H}_{CC} \succeq \lambda I_{|C|}$ with $\lambda > 0$,

$$\|\mathcal{H}_{CC}^{-1}\|_2 \le \lambda^{-1}.$$

It follows that

$$\begin{aligned}\|z_C\|_2 &\le \|\mathcal{H}_{CC}^{-1}\|_2 \left(\|a_C\|_2 + \|e_C\|_2 + \|\mathcal{H}_{CR}\|_2\|z_R\|_2\right) \\ &\le \frac{\varepsilon_C + \tau_C}{\lambda} + \rho\|z_R\|_2 = B_C.\end{aligned}$$

This proves Equation equation 13. For every $j \in C$,

$$|\Delta_j| = \frac{|z_j|}{\sqrt{E_j}} \le \frac{\|z_C\|_2}{\sqrt{E_j}} \le \frac{B_C}{\sqrt{E_j}},$$

which establishes Equation equation 14. □

Selective adaptation and representation stability. Suppose $\|z_R\|_2 > 0$ and

$$\frac{\varepsilon_C + \tau_C}{\lambda} \le \alpha\|z_R\|_2, \qquad \alpha \ge 0, \quad q := \alpha + \rho < 1.$$

Then Proposition 1 gives

$$\|z_C\|_2 \le q\|z_R\|_2 < \|z_R\|_2. \tag{32}$$

The selected group therefore undergoes smaller aggregate energy-weighted log-gain movement.

Combining the same proposition with Equation equation 30 gives the direct representation bound

$$\|\Delta H_C^{\mathrm{direct}}\|_F \le e^{\delta_C}\left(\frac{\varepsilon_C + \tau_C}{\lambda} + \rho\|z_R\|_2\right). \tag{33}$$

Moreover,

$$\frac{\|\Delta H_C^{\mathrm{direct}}\|_F}{\|\Delta H_R^{\mathrm{direct}}\|_F} \le e^{\delta_C + \delta_R} q. \tag{34}$$

Thus, $e^{\delta_C+\delta_R}q < 1$ establishes smaller direct representation displacement in the selected group. At an interior stationary endpoint, $\nabla J(\theta_1) = 0$ and $\tau_C = 0$.

The fixed-batch GRPO reward surrogate. Let a rollout batch contain B prompts x_b and $G \ge 2$ completions y_{bi} per prompt, with completion lengths T_{bi} and rewards r_{bi}. Define the group mean reward and normalized advantages by

$$\bar{r}_b = \frac{1}{G}\sum_{i=1}^{G} r_{bi}, \qquad A_{bi} = \frac{r_{bi} - \bar{r}_b}{s_{r,b} + \epsilon_{\mathrm{adv}}},$$

where $s_{r,b}$ is the group reward standard deviation used by the algorithm and $\epsilon_{\mathrm{adv}} > 0$ is its stabilizing constant.

For completion token t, define the policy ratio

$$\varrho_{bit}(\theta) = \frac{\pi_\theta(y_{bi,t} \mid x_b, y_{bi,<t})}{\pi_{\theta_0}(y_{bi,t} \mid x_b, y_{bi,<t})}.$$

With clipping width $\epsilon_{\mathrm{clip}} \in (0,1)$, the length-normalized reward surrogate is

$$\begin{aligned}U(\theta) = \frac{1}{BG}\sum_{b=1}^{B}\sum_{i=1}^{G}\frac{1}{T_{bi}}\sum_{t=1}^{T_{bi}}\min\Big\{&\varrho_{bit}(\theta)A_{bi}, \\ &\mathrm{clip}\big(\varrho_{bit}(\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}}\big)A_{bi}\Big\}.\end{aligned} \tag{35}$$

All sampled tokens and advantages remain fixed when differentiating this surrogate.

At θ_0, every ratio equals one. Define the length-normalized completion score

$$s_{bi} = \frac{1}{T_{bi}} \nabla_\theta \log \pi_\theta(y_{bi} \mid x_b)\Big|_{\theta=\theta_0}, \qquad \bar{s}_b = \frac{1}{G}\sum_i s_{bi}.$$

The reward gradient contributed by prompt b is

$$\begin{aligned} g_b^{\mathrm{rew}} &= \frac{1}{G}\sum_i A_{bi} s_{bi} \\ &= \frac{1}{G(s_{r,b}+\epsilon_{\mathrm{adv}})}\sum_i (r_{bi}-\bar{r}_b)(s_{bi}-\bar{s}_b). \end{aligned} \tag{36}$$

The second equality follows from $\sum_i (r_{bi}-\bar{r}_b) = 0$. Consequently, $\nabla U(\theta_0) = B^{-1}\sum_b g_b^{\mathrm{rew}}$.

Bounding the normalized reward signal. Define empirical variances with denominator G:

$$\sigma_{r,b}^2 = \frac{1}{G}\sum_i (r_{bi}-\bar{r}_b)^2, \qquad \sigma_{s,b,j}^2 = \frac{1}{G}\sum_i (s_{bi,j}-\bar{s}_{b,j})^2.$$

Applying Cauchy–Schwarz to Equation equation 36 gives

$$|g_{b,j}^{\mathrm{rew}}| \le \frac{\sigma_{r,b}}{s_{r,b}+\epsilon_{\mathrm{adv}}}\sigma_{s,b,j}.$$

Squaring, dividing by E_j, and summing over $j \in C$ yields

$$\|D_C^{-1/2} g_{b,C}^{\mathrm{rew}}\|_2 \le \frac{\sigma_{r,b}}{s_{r,b}+\epsilon_{\mathrm{adv}}} \left(\sum_{j\in C} \frac{\sigma_{s,b,j}^2}{E_j}\right)^{1/2}.$$

Therefore, defining

$$\varepsilon_C^{\mathrm{rew}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sigma_{r,b}}{s_{r,b}+\epsilon_{\mathrm{adv}}} \left(\sum_{j\in C} \frac{\sigma_{s,b,j}^2}{E_j}\right)^{1/2}, \tag{37}$$

the triangle inequality gives

$$\|D_C^{-1/2}\nabla_C U(\theta_0)\|_2 \le \varepsilon_C^{\mathrm{rew}}.$$

Since $J = U - \beta K$,

$$\varepsilon_C \le \varepsilon_C^{\mathrm{rew}} + \beta \|D_C^{-1/2}\nabla_C K(\theta_0)\|_2. \tag{38}$$

This expresses the initial normalized optimization signal in terms of reward-correlated score variation and the reference-policy gradient. A score coordinate shared across all completions in a group has $\sigma_{s,b,j} = 0$ and contributes zero reward gradient.

Sufficient curvature conditions. The curvature constants in the stability theorem can be obtained from the reward and KL terms separately. For $s \in [0,1]$, define

$$A_K(s) = D^{-1/2}\nabla^2 K(\theta_0 + s\Delta) D^{-1/2}, \qquad A_U(s) = D^{-1/2}\nabla^2 U(\theta_0 + s\Delta) D^{-1/2}.$$

Suppose that throughout the segment,

$$\begin{aligned} [A_K(s)]_{CC} \succeq \kappa I_{|C|}, \qquad [A_U(s)]_{CC} \preceq \gamma I_{|C|}, \\ \|[A_K(s)]_{CR}\|_2 \le \kappa_\times, \qquad \|[A_U(s)]_{CR}\|_2 \le \gamma_\times, \end{aligned} \tag{39}$$

where $\kappa > 0$, $\gamma, \kappa_\times, \gamma_\times \ge 0$, and $\beta\kappa > \gamma$. Because

$$\mathcal{H} = \int_0^1 \big(\beta A_K(s) - A_U(s)\big)\, ds,$$

integration and the triangle inequality give

$$\mathcal{H}_{CC} \succeq (\beta\kappa - \gamma) I_{|C|}, \qquad \|\mathcal{H}_{CR}\|_2 \le \beta\kappa_\times + \gamma_\times.$$

Thus, Proposition 1 applies with

$$\lambda = \beta\kappa - \gamma, \qquad \rho = \frac{\beta\kappa_\times + \gamma_\times}{\beta\kappa - \gamma}. \tag{40}$$

These constants quantify the restoring curvature and cross-group coupling of the complete KL-regularized objective.

Curvature at the reference policy. For a fixed prompt distribution d and a smooth, full-support policy on a finite completion space, consider the sequence-level KL

$$K_{\mathrm{seq}}(\theta) = \mathbb{E}_{x\sim d} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\big).$$

Let θ_{ref} satisfy $\pi_{\theta_{\mathrm{ref}}} = \pi_{\mathrm{ref}}$. Normalization of π_θ gives

$$\nabla K_{\mathrm{seq}}(\theta) = \mathbb{E}_{x\sim d} \sum_y \nabla \pi_\theta(y \mid x) \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}.$$

At θ_{ref}, the logarithm vanishes, so $\nabla K_{\mathrm{seq}}(\theta_{\mathrm{ref}}) = 0$. Differentiating once more yields

$$\nabla^2 K_{\mathrm{seq}}(\theta_{\mathrm{ref}}) = \mathbb{E}_{x\sim d,\, y\sim\pi_{\mathrm{ref}}}[q(x,y)q(x,y)^\top], \tag{41}$$

where

$$q(x,y) = \nabla_\theta \log \pi_\theta(y \mid x)|_{\theta=\theta_{\mathrm{ref}}}.$$

Hence the reference-policy curvature is the Fisher matrix.

The sampled token-level k_3 penalty has an analogous empirical identity. For fixed contexts ξ_n, tokens a_n, and positive weights ω_n summing to one, define

$$\chi_n(\theta) = \frac{\pi_{\mathrm{ref}}(a_n \mid \xi_n)}{\pi_\theta(a_n \mid \xi_n)}, \qquad K_{k_3}(\theta) = \sum_n \omega_n \big(\chi_n(\theta) - \log \chi_n(\theta) - 1\big).$$

Writing $q_n(\theta) = \nabla_\theta \log \pi_\theta(a_n \mid \xi_n)$, differentiation gives

$$\nabla K_{k_3}(\theta) = \sum_n \omega_n \big(1 - \chi_n(\theta)\big) q_n(\theta).$$

At the reference-policy anchor, $\chi_n = 1$, and therefore

$$\begin{aligned} \nabla K_{k_3}(\theta_{\mathrm{ref}}) &= 0, \\ \nabla^2 K_{k_3}(\theta_{\mathrm{ref}}) &= \sum_n \omega_n q_n(\theta_{\mathrm{ref}}) q_n(\theta_{\mathrm{ref}})^\top. \end{aligned} \tag{42}$$

Energy normalization transforms these Fisher and empirical Gram matrices into Gram matrices of energy-normalized policy scores.

For either penalty, when $\theta_0 = \theta_{\mathrm{ref}}$, Equation equation 38 becomes

$$\varepsilon_C \le \varepsilon_C^{\mathrm{rew}}.$$

Together, the reward-signal bound, the segment-wise curvature conditions, and the direct gate geometry establish an explicit criterion for selective adaptation in the energy-weighted representation metric.

A.4 Channel Subspace Metrics and Calibrations

This section provides formal proofs for the empirical subspace metrics evaluated in Section 3.

Proposition 6 (Overlap as Mean-Squared Principal Cosine)**.** *Let $U_g^+ \in \mathbb{R}^{D\times r_g^+}$ and $U_g^- \in \mathbb{R}^{D\times r_g^-}$ denote orthonormal bases for the correct and wrong representation subspaces in channel group $g \in \{C, R\}$, with $k = \min(r_g^+, r_g^-)$. The overlap metric defined in Equation (2) of the main text satisfies:*

$$\mathcal{O}_g^2 = \frac{\|(U_g^+)^\top U_g^-\|_F^2}{k} = \frac{1}{k}\sum_{j=1}^k \cos^2\phi_j = \frac{\mathrm{tr}(P_g^+ P_g^-)}{k}, \qquad \mathcal{O}_g \in [0,1], \tag{43}$$

where $\phi_1 \le \cdots \le \phi_k$ are the principal angles between the two subspaces, and $P_g^\pm = U_g^\pm (U_g^\pm)^\top$.

Proof. By the Jordan–Wielandt theorem, the singular values of $(U_g^+)^\top U_g^-$ are exactly the principal cosines $\sigma_j = \cos\phi_j \in [0,1]$ for $j \in \{1,\dots,k\}$. Thus:

$$\|(U_g^+)^\top U_g^-\|_F^2 = \sum_{j=1}^k \sigma_j^2 = \sum_{j=1}^k \cos^2\phi_j. \tag{44}$$

Using the cyclic commutativity of the trace:

$$\mathrm{tr}(P_g^+ P_g^-) = \mathrm{tr}\big(U_g^+ (U_g^+)^\top U_g^- (U_g^-)^\top\big) = \mathrm{tr}\big((U_g^+)^\top U_g^- (U_g^-)^\top U_g^+\big) = \|(U_g^+)^\top U_g^-\|_F^2. \quad (45)$$

Dividing by k yields the claim. □

Proposition 7 (Expected Overlap Under Isotropic Null Calibration)**.** *Let U_g^+ and U_g^- be independent, uniformly distributed random subspaces on the Grassmannian $\mathbf{Gr}(r_g^\pm, D)$. Then:*

$$\mathbb{E}\big[\mathcal{O}_g^2\big] = \frac{\max(r_g^+, r_g^-)}{D}. \quad (46)$$

Proof. By rotational invariance of the Haar measure, $\mathbb{E}[P_g^-] = \frac{r_g^-}{D} I_D$. Conditioning on P_g^+ and using independence:

$$\mathbb{E}\big[\mathrm{tr}(P_g^+ P_g^-) \mid P_g^+\big] = \mathrm{tr}\left(P_g^+ \frac{r_g^-}{D} I_D\right) = \frac{r_g^-}{D}\,\mathrm{tr}(P_g^+) = \frac{r_g^+ r_g^-}{D}. \quad (47)$$

Dividing by $k = \min(r_g^+, r_g^-)$ yields $\frac{r_g^+ r_g^-}{D \min(r_g^+, r_g^-)} = \frac{\max(r_g^+, r_g^-)}{D}$. □

Remark (Null Calibration Context). Proposition 7 demonstrates that when the ambient channel dimension is large ($D \gg r$), independent random subspaces yield an expected overlap near zero. The empirical overlap observed in Figure 1(a) ($\mathcal{O}_C \approx 1.0$) thus reflects mathematically profound shared structure between correct and wrong rollouts.

Proposition 8 (Projector Distance Equals Square Root of Symmetric Difference)**.** *Let $C^{(t)}, C^{(\mathrm{pre})} \subseteq \{1, \ldots, d\}$ be channel sets with diagonal projectors $P_C^{(t)} = \mathrm{diag}(\mathbf{1}_{C^{(t)}})$. The Frobenius distance between projectors satisfies:*

$$\|P_C^{(t)} - P_C^{(\mathrm{pre})}\|_F = \sqrt{|C^{(t)} \triangle C^{(\mathrm{pre})}|}, \quad (48)$$

where $\triangle$ denotes symmetric set difference.

Proof. $P_C^{(t)} - P_C^{(\mathrm{pre})}$ is a diagonal matrix whose j-th entry is non-zero if and only if index j belongs to one set but not the other:

$$(P_C^{(t)} - P_C^{(\mathrm{pre})})_{jj} = \begin{cases} +1, & j \in C^{(t)} \setminus C^{(\mathrm{pre})}, \\ -1, & j \in C^{(\mathrm{pre})} \setminus C^{(t)}, \\ 0, & \text{otherwise.} \end{cases} \quad (49)$$

Squaring the diagonal entries and summing across all $j \in \{1, \ldots, d\}$ yields $|C^{(t)} \triangle C^{(\mathrm{pre})}|$. Taking the square root proves Equation (3) of the main text. □

Proposition 9 (Exact Energy Decomposition of Baseline-Relative Drift)**.** *Let the channels $\{1, \ldots, d\}$ be partitioned into orthogonal coherence channels C and reasoning channels R, with projectors $P_C^{(0)}$ and $P_R^{(0)} = I - P_C^{(0)}$. For any activation displacement $\Delta H = H^{(1)} - H^{(0)}$, define raw drift F_g, normalized percentage drift A_g, and baseline energy weight w_g:*

$$F_g = \|\Delta H P_g^{(0)}\|_F, \qquad A_g = 100 \times \frac{F_g}{\|H^{(0)} P_g^{(0)}\|_F}, \qquad w_g = \frac{\|H^{(0)} P_g^{(0)}\|_F^2}{\|H^{(0)}\|_F^2}, \quad g \in \{C, R\}. \quad (50)$$

Then the total normalized activation drift satisfies:

$$\frac{\|\Delta H\|_F^2}{\|H^{(0)}\|_F^2} = w_C \left(\frac{A_C}{100}\right)^2 + w_R \left(\frac{A_R}{100}\right)^2. \quad (51)$$

Proof. Because $P_C^{(0)}$ and $P_R^{(0)}$ project onto orthogonal coordinate subsets, the squared Frobenius norms decompose additively:

$$\|\Delta H\|_F^2 = \|\Delta H P_C^{(0)}\|_F^2 + \|\Delta H P_R^{(0)}\|_F^2 = F_C^2 + F_R^2, \quad (52)$$

$$\|H^{(0)}\|_F^2 = \|H^{(0)} P_C^{(0)}\|_F^2 + \|H^{(0)} P_R^{(0)}\|_F^2. \quad (53)$$

Dividing both sides of the displacement sum by $\|H^{(0)}\|_F^2$:

$$\frac{\|\Delta H\|_F^2}{\|H^{(0)}\|_F^2} = \frac{\|H^{(0)} P_C^{(0)}\|_F^2}{\|H^{(0)}\|_F^2} \left(\frac{F_C}{\|H^{(0)} P_C^{(0)}\|_F} \right)^2 + \frac{\|H^{(0)} P_R^{(0)}\|_F^2}{\|H^{(0)}\|_F^2} \left(\frac{F_R}{\|H^{(0)} P_R^{(0)}\|_F} \right)^2 . \tag{54}$$

Substituting w_g and $A_g/100$ establishes the decomposition, confirming the validity of Equation (4). □