RANKING IS NOT CONTROL: REFERENCE-GROUNDED STEERING FOR DISTINCT BEHAVIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation steering provides lightweight control over frozen language models, and many optimization-based methods learn steering vectors from pairwise preferences. We argue that this formulation confuses ranking with control: objectives on ranking determine relative preference, but not the direction or selectivity of behavioral change. Unconstrained neutral or irrelevant outcomes mean that pairwise advancement does not guarantee intended uplift: the same positive margin can reflect intended uplift, joint suppression, or broad co-amplification. Thus, we formulate representation steering as reference-grounded constrained behavioral control and instantiate it as Reference-Grounded Steering (RGS), a soft-penalty objective that promotes intended uplift, penalizes grounding and separation violations, limits competing-behavior leakage, and regularizes intervention energy. Together, these terms constrain the behavioral movements that pairwise preferences leave unspecified. Controlled experiments validate the predicted failure mode: pairwise-only objectives attain large positive relative margins while decreasing support for the intended behavior, and additional endpoint flexibility amplifies this failure when the corresponding behavioral constraints are absent. Once reference-grounded constraints are introduced, the same flexibility becomes beneficial. Across nine models from three families, RGS achieves the strongest bidirectional control, ordered bracketing of the unsteered policy, and control against the prompt-designated orientation across held-out behavior-pair combinations that were never jointly trained or calibrated. Along interpolated endpoint chords, RGS also provides a wider control range with fewer and smaller local reversals than the baselines. These results suggest that effective representation steering requires specifying not only which behavior should rank higher, but the structured pattern of behavioral change relative to the unsteered model.

1 INTRODUCTION

Representation steering offers a lightweight control over frozen language models by modifying internal activations at inference time (Subramani et al., 2022; Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023; Wu et al., 2024). Recent optimization-based methods, such as BiPO (Cao et al., 2024) and RePS (Wu et al., 2025b), learn such interventions from preferences between candidate responses, making pairwise objectives a natural interface for training steering vectors.

We argue that ranking is not behavioral control. A relative objective specifies which behavior should score higher, but selective control also depends on how behavior moves relative to the unsteered model. The same positive margin can arise from intended uplift, but also from alternative suppression, unequal joint suppression, or broad co-amplification, while responses outside the pair remain unconstrained. Relative improvement alone therefore cannot establish intended uplift or selectivity. Figure 1 illustrates this distinction.

We introduce Reference-Grounded Steering (RGS) as soft optimization of reference-relative response-score conditions, using independently learned endpoints. Its zero-anchored objective promotes intended uplift over alternative and native-response movement and separately limits alternative uplift. All displacements compare fixed responses against the same frozen, unsteered model.

We evaluate our steering method in human-value conflicts, which provides a demanding testbed for bidirectional, selective, and reusable controls across changing value priorities. In these settings,

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.

the desirable response can depend on the person, situation, and culture, so a single ordering of values may not suit every context (Schwartz, 1992; Santurkar et al., 2023; Sorensen et al., 2024; Zhang et al., 2026). Recent benchmarks reflect this diversity by evaluating value priorities across cultures and decision scenarios (Ren et al., 2024; Yao et al., 2024; Zhao et al., 2024; Liu et al., 2025; Tang & Yeh, 2026). A deployed model should therefore allow its expressed priority between competing values to adapt to the user and context. Doing so without retraining the full policy motivates lightweight behavioral control.

We test nine models on conflicts among the ten Schwartz values, testing transfer to held-out value-pair combinations without retuning. RGS improves selective bidirectional control over BiPO and RePS on every model. Its nine-model mean quality loss is larger than BiPO's and smaller than RePS's. Matched ablations connect these control gains to the objective's components.

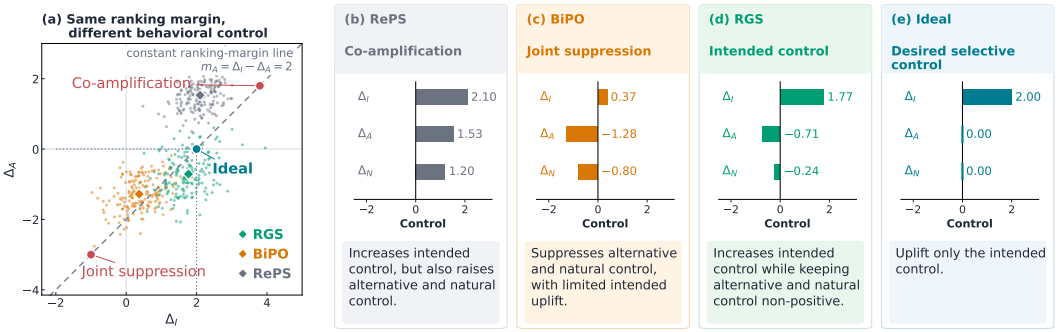


Figure 1: Ranking is not control. RGS aligns steering with the desired selective-control regime.

Our contributions are threefold: **Reference-grounded steering.** We identify pair-wise steering underdetermines the direction and selectivity of representation, formulate appropriate steering as control in reference-relative behavioral co-ordinates, and introduce RGS with independent endpoints and a zero-anchored multi-competitor objective. **Matched mechanism evidence.** Matched ablations distinguish zero anchoring, intended-score enhancement, leakage control, and endpoint parameterization. Under full $0/A/N$ grounding, enhancement and leakage control improve intended scores and quality while largely preserving the separate-endpoint BPC advantage. **Frozen pair-transfer evaluation.** We select global operating points only on five selection pairs and evaluate without retuning on 40 held-out transfer pairs. We report control together with response quality and unclear/refusal outcomes.

2 Related Work

Preference alignment and plural objectives. Pairwise feedback underlies RLHF and direct policy objectives such as DPO (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023); later work changes the preference objective or its coupling to supervised learning (Gheshlaghi Azar et al., 2024; Ethayarajh et al., 2024; Meng et al., 2024; Hong et al., 2024). RGS addresses a distinct ambiguity: even with correct pair labels, a relative steering margin does not locate either response relative to the unsteered model. Plural-alignment methods instead select or combine policy objectives (Ramé et al., 2023; Zhong et al., 2024; Zhou et al., 2024; Guo et al., 2024; Wang et al., 2024; Bakker et al., 2022); RGS studies the behavioral coordinates needed to move a frozen model selectively between endpoints.

Representation steering. Steering modifies hidden states of a frozen model. Existing methods extract directions from target or contrastive examples, probes, or learned interventions (Subramani et al., 2022; Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023; Wu et al., 2024). Reliability and selectivity remain central concerns: interventions can mediate refusal, vary with inputs, and trade control against capability (Arditi et al., 2024; Lee et al., 2025; Tan et al., 2024; Wu et al., 2025a). RGS complements this work by defining selectivity through response-score movement relative to a frozen reference. BiPO and RePS are the closest optimization-based baselines. BiPO learns a single signed axis for a behavior pair (Cao et al., 2024); RePS uses an asymmetric objective over

conditioned and unconditioned responses, and its steering-vector instantiation combines positive addition with projection nulling (Wu et al., 2025b).

Human values and behavioral evaluation Schwartz organizes ten values by motivational compatibility and conflict (Schwartz, 1992); LLM studies further measure value orientations, social decisions, and cultural variation (Ren et al., 2024; Yao et al., 2024; Liu et al., 2025; Zhao et al., 2024). We use CoopValue's situated conflicts to evaluate endpoint control on held-out pair combinations (Tang & Yeh, 2026). ConVA steers individual Schwartz values, whereas RGS controls conflicting decision endpoints (Jin et al., 2025); steering tokens study unseen combinations on an input-token rather than activation-vector control surface (Radevski et al., 2026). We use a method-blind LLM judge for open-ended evaluation (Liu et al., 2023; Zheng et al., 2023).

3 Reference-Grounded Steering

Pairwise improvement does not determine how either response changes relative to the unsteered model: the same margin gain can arise from intended uplift, alternative suppression, or unequal joint shifts, while responses outside the pair remain unconstrained. We introduce **Reference-Grounded Steering** (RGS), which tracks score changes for intended, alternative, and model-native responses against the unsteered model and optimizes a selective-control specification. To avoid repeated generation and behavioral evaluation during training, we use teacher-forced scores on fixed response pools as a differentiable surrogate. The resulting behavioral changes are assessed separately through free generation. We instantiate this framework on value-conditioned behavior pairs.

3.1 Reference-Grounded Coordinates

Let p_θ be a frozen autoregressive language model and B_1, B_2 two behavior patterns. For each question x_i, we construct fixed response pools $\mathcal{Y}_{1,i}, \mathcal{Y}_{2,i}$ conditioned on B_1, B_2, and a native pool $\mathcal{Y}_{N,i}$ generated without behavior conditioning. The dataset is $\mathcal{D} = \{(x_i, \mathcal{Y}_{1,i}, \mathcal{Y}_{2,i}, \mathcal{Y}_{N,i})\}_{i=1}^n$. A sampled record $z_i = (x_i, y_{1,i}, y_{2,i}, y_{N,i})$ draws one response from each pool; $\mathbb{E}_\mathcal{D}$ averages over these draws with equal source weights. Native responses need not be behaviorally neutral.

For a response $y = (y_1, \ldots, y_L)$ and activation offset $\xi \in \mathbb{R}^D$, define the normalized discounted teacher-forced score

$$S_\xi(y \mid x) = \sum_{t=1}^{L} \omega_t \log p_\theta(y_t \mid x, y_{<t}; \xi), \qquad \omega_t = \frac{t^{-\gamma}}{\sum_{j=1}^{L} j^{-\gamma}}, \quad \gamma \geq 0. \tag{1}$$

Thus S_ξ is a response-level log-likelihood proxy, with earlier tokens receiving greater weight when $\gamma > 0$. For each direction $d \in \{1, 2\}$, RGS learns an endpoint $\boldsymbol{u}_d \in \mathbb{R}^D$ and intervenes at layer l by

$$h'_{l,t} = h_{l,t} + \alpha \boldsymbol{u}_d, \tag{2}$$

where the offset is applied at positions whose logits predict answer tokens, beginning at the last prompt position. Training fixes $\alpha = 1$; inference varies α. The endpoints are independently parameterized, without requiring $\boldsymbol{u}_2 = -\boldsymbol{u}_1$, allowing asymmetric interventions for the two directions.

For direction B_1, we assign $(y_I^{(1)}, y_A^{(1)}, y_N^{(1)}) = (y_1, y_2, y_N)$; for direction B_2, we assign $(y_I^{(2)}, y_A^{(2)}, y_N^{(2)}) = (y_2, y_1, y_N)$. Here I denotes the intended response, A the alternative response, and N the native response. Define the reference-relative score displacement as

$$\Delta_X^{(d)}(\alpha) := S_{\alpha \boldsymbol{u}_d}(y_X^{(d)} \mid x) - S_0(y_X^{(d)} \mid x) = \log \frac{\exp S_{\alpha \boldsymbol{u}_d}(y_X^{(d)} \mid x)}{\exp S_0(y_X^{(d)} \mid x)}, \qquad X \in \{I, A, N\}, \tag{3}$$

This is a weighted log-probability change for a fixed response, equivalently a log ratio of likelihood proxies rather than full sequence probabilities. Positive Δ_I indicates intended uplift; negative Δ_A indicates alternative suppression. The unsteered score S_0 grounds every coordinate, whereas y_N supplies an additional response to monitor. Its displacement Δ_N does not measure aggregate probability mass outside the pair.

An objective depending only on $m_A = \Delta_I - \Delta_A$ is invariant to a common shift of Δ_I and Δ_A and leaves Δ_N unspecified. For example, $(\Delta_I, \Delta_A) = (1, 0), (2, 1), (-1, -2)$ all yield $m_A =$

1, despite different absolute movements. RGS therefore controls the full coordinate triple $\Delta = (\Delta_I, \Delta_A, \Delta_N)$. Below, we suppress record and direction indices and use $\alpha = 1$ unless otherwise stated.

3.2 Selective-Control Specification

For desired margin $\rho > 0$ and allowable alternative uplift $\epsilon \geq 0$, RGS specifies the following per-record control conditions:

$$\Delta_I \geq \rho, \quad \text{(G: intended uplift)}, \quad \Delta_I - \Delta_A \geq \rho, \quad (\mathrm{S}_A\text{: competition separation}), \tag{4}$$

$$\Delta_A \leq \epsilon, \quad \text{(L: alternative uplift bound)}. \quad \Delta_I - \Delta_N \geq \rho, \quad (\mathrm{S}_N\text{: native separation}), \tag{5}$$

(G) requires positive intended movement. (S_A) and (S_N) require it to exceed alternative and native movement, respectively. These relative conditions still allow joint amplification, so (L) separately caps alternative uplift. Native separation does not require preserving the native score or the broader response distribution.

For one endpoint, the ideal empirical control problem requires the conditions on every sampled record:

$$\max_{\boldsymbol{u}_d} \quad \mathbb{E}_{\mathcal{D}} \Delta_I^{(d)}, \qquad \text{s.t. (G), } (\mathrm{S}_A), (\mathrm{S}_N), \text{ and (L).} \tag{6}$$

These conditions define a target specification. Since $S_\xi \leq 0$, intended uplift satisfies $\Delta_I \leq -S_0(y_I \mid x)$, making (G) infeasible whenever $-S_0(y_I \mid x) < \rho$. Shared endpoints may further limit joint feasibility, motivating a soft surrogate.

3.3 Conservative Surrogate Objective

The violations of the three margin conditions are

$$v_0 = \rho - \Delta_I, \qquad v_A = \rho - \Delta_I + \Delta_A, \qquad v_N = \rho - \Delta_I + \Delta_N. \tag{7}$$

Their worst violation is $M = \max\{v_0, v_A, v_N\} = \rho - \Delta_I + \max\{0, \Delta_A, \Delta_N\}$. The zero anchor requires intended uplift even when both competing responses are suppressed; suppressing competitors alone cannot satisfy (G).

To obtain a smooth conservative surrogate, we define $C_\tau := \mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$, i.e., $C_\tau = \tau \log\left(1 + e^{\Delta_A/\tau} + e^{\Delta_N/\tau}\right)$ where $\tau > 0$, and $m_\tau = \Delta_I - C_\tau, M_\tau = \rho - m_\tau$. Since

$$\max\{0, \Delta_A, \Delta_N\} \leq C_\tau \leq \max\{0, \Delta_A, \Delta_N\} + \tau \log 3, \tag{8}$$

we have $M \leq M_\tau \leq M + \tau \log 3$. Thus M_τ is an upper approximation to the worst margin violation M, and $M_\tau \leq 0$ implies satisfaction of (G), (S_A), and (S_N).

We apply the softplus penalty $\phi_\beta(M_\tau) = \log(1 + e^{\beta M_\tau}),\ \beta > 0$, whose β controls both sharpness and gradient scale, and use a hinge penalty for alternative uplift: $[\Delta_A - \epsilon]_+$. For record z_i and direction d,

$$\ell_d(z_i) = -\Delta_{I,i}^{(d)} + \lambda_1 \phi_\beta(M_{\tau,i}^{(d)}) + \lambda_2 [\Delta_{A,i}^{(d)} - \epsilon]_+. \tag{9}$$

The resulting objective for the two independently parameterized endpoints with symmetric directional losses is

$$\widehat{\mathcal{L}} = \mathbb{E}_{\mathcal{D}} \left[\ell_1 + \ell_2\right] + \lambda_3 \left(\|\boldsymbol{u}_1\|^2 + \|\boldsymbol{u}_2\|^2\right). \tag{10}$$

The explicit $-\Delta_I$ term continues to reward uplift as the margin penalty weakens beyond the target margin. The softplus penalizes the conservative worst margin violation, the hinge limits alternative uplift, and the squared norms regularize intervention energy. The λ's are fixed penalty weights, not Lagrange multipliers; minimizing this objective does not guarantee per-record constraint satisfaction.

3.4 Training and Inference

For each source, we generate and freeze three responses per family and cache their unsteered scores. Each epoch visits every source once and samples one response from each family. We compute the two-direction loss on each sampled triplet before averaging, and update only $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ with

minibatch gradients through the frozen model. After each AdamW step, each endpoint is projected onto $\|\boldsymbol{u}_d\|_2 \leq 30$. Teacher forcing uses $\alpha = 1$ and the injection scope in Equation 2.

At inference, the target direction d selects the offset $\alpha \boldsymbol{u}_d$. Training at $\alpha = 1$ does not imply monotonic control at other strengths. Layer and strength are selected using the protocol in Section 4, which also evaluates freely generated behavior using bidirectional preference contrast (BPC). Data construction and optimization details are given there and in the accompanying appendices.

4 EXPERIMENTS

We test how reference grounding, objective terms, endpoint parameterization, and training-pair structure affect control, then evaluate transfer to new value pairs across nine models and behavior between steering endpoints.

4.1 EXPERIMENTAL SETUP

Models and training data. We compare our RGS method with RePS, and BiPO on nine Qwen3, Ministral-3, and Llama models using text-only inputs. We use five training value pairs covering the ten Schwartz values, with shared scenarios and model-generated response pools across methods. Each model generates value-conditioned and natural responses, and each method learns separate steering parameters for the five pairs at prespecified middle-layer candidates while retaining its native objective and intervention mechanism. Data construction, model specifications, and training details are given in Appendices B and C.

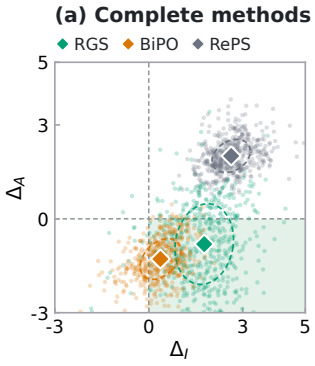
Evaluation and transfer. CoopValue provides evaluation scenarios for all ($\binom{10}{2} = 45$) unordered value pairs. We use the 5 training value pairs for configuration selection; their 196 scenarios are used only to select a steering configuration (layer and strength). The remaining 40 pairs (transfer value pairs), comprising 1582 scenarios, are held out for transfer evaluation. For each method on each model, we select one steering configuration on all training pairs and keep it fixed for transfer evaluation.

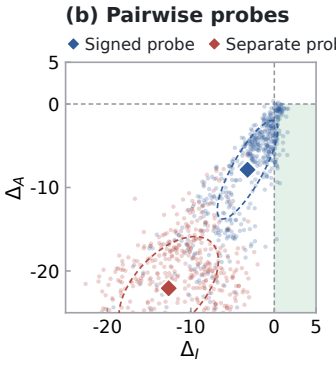
Generation and metrics. Greedy decoding produces unsteered and bidirectionally steered responses. A method-blind Deepseek-v4-flash judge scores value support and quality. Bidirectional Preference Contrast (BPC) is the primary free-generation outcome; we also report Order-preserving rate, Target-value Induction Rate, and quality (Appendices C and D). These metrics do not establish absolute intended-support uplift over the unsteered output or satisfaction of the score-space specification.

Configuration selection and diagnostics. On training pairs, we maximize BPC subject to $q \geq -0.15$, breaking ties by Order-preserving rate and quality change. The configuration is fixed for transfer and interpolation. Teacher-forced validation displacements diagnose fixed-response score changes (Appendix F).

4.2 PAIRWISE SEPARATION DOES NOT GUARANTEE INTENDED UPLIFT

A positive pairwise margin need not increase intended-response scores. Figure 2(a) overlays the complete methods: on average, BiPO combines modest intended uplift with alternative suppression, RePS raises both scores, and RGS combines positive mean intended-response score displacement with negative mean alternative-response score displacement. The two matched RGS-pipeline probes in Figure 2(b) isolate a score-based loss depending only on $\Delta_I - \Delta_A$, with enhancement and alternative leakage disabled. The signed-axis probe attains a mean margin of 4.663 with $\Delta_I = -3.208$; separate endpoints yield 9.418 with $\Delta_I = -12.651$. At this inference strength, both probes exhibit positive margins alongside joint score suppression (Appendix F.4.1). Panel (c) reports mean native-response score displacement and the **negative-intended rate**, the fraction of records with $\Delta_I < 0$.





(c) Background and uplift

500 directional records per row

	Mean Δ_N	$\Delta_I < 0$
RGS	-0.242	3.4%
BiPO	-0.803	25.8%
RePS	+1.583	0.0%
Signed probe	-5.952	83.0%
Separate probe	-18.467	100.0%

Figure 2: Qwen3-4B Validation displacements. (a) Complete methods at their own selected steering configurations; (b) pairwise probes at the RGS configuration (layer 20, inference strength 1.5; trained at strength 1); (c) mean native-response score displacement and negative-intended rates. Each configuration contains $5 \times 50 \times 2 = 500$ source–direction records under shared teacher-forced scoring. Colors identify configurations; diamonds mark means. Dashed ellipses show one-standard-deviation covariance contours under a Gaussian approximation; they are descriptive rather than regression fits or confidence intervals. Dashed coordinate lines mark zero displacement; shading marks $\Delta_I > 0$, $\Delta_A \leq 0$. Panels (a) and (b) use different ranges, with equal horizontal and vertical units within each panel.

4.3 Ablation Studies

Table 1 brings together alternative-set, enhancement/alternative leakage, and endpoint-parameterization ablations at the same Qwen3-4B steering configuration. The matched designs and component-results scope are specified in Appendix F.2 and Appendix F.4.2. Here, E enables enhancement, the intended-uplift term $-\Delta_I$, and L enables the alternative-uplift penalty $\lambda_2[\Delta_A - \epsilon]_+$. The margin penalty and energy regularization are retained in every configuration. For each record, we define $m_{\text{worst}} = \min\{\Delta_I, \Delta_I - \Delta_A, \Delta_I - \Delta_N\}$. We first take this minimum per record, then average records within each pair and equally across the five pairs. The result need not equal the minimum computed from the three tabulated mean displacements. Q_s is absolute steered quality (1–5), averaging the judge's relevance, coherence, and specificity scores per response, then both steering directions and prompt orientations within each source, sources within each pair, and the 40 transfer pairs equally.

Reference grounding and alternative coverage. With separate endpoints and both enhancement and alternative leakage disabled, we add alternative-response terms to the same log-sum-exp (LSE) construction. Since $\text{LSE}_\tau(\Delta_A) = \Delta_A$, the first transition adds the zero-uplift reference, making mean intended-response score displacement positive and improving transfer-pair BPC by $0.233\,[0.218, 0.248]$. Adding natural-response competition further increases BPC by $0.036\,[0.024, 0.048]$, while mean intended-response score displacement falls from 1.307 to 0.880 and mean steered quality decreases by 0.308. Grounding changes how separation is achieved; natural-response competition also introduces a control–quality tradeoff.

Enhancement and alternative leakage. For separate endpoints, a 2×2 comparison fixes the full $0/A/N$ alternative set and varies the two additional terms. Enhancement alone raises mean Δ_I from 0.880 to 2.279, but also raises alternative and native scores, reducing BPC from .623 to .454. Alternative leakage alone does not improve intended uplift, worst-case target margin, or BPC over the configuration without enhancement or alternative leakage. Adding alternative leakage to enhancement lowers mean Δ_A from +1.706 to -0.809 and restores BPC to .624. Relative to the configuration without either term, the full combination increases mean Δ_I by 0.891, m_{worst} by 0.520, and Q_s by 0.290, while BPC remains nearly unchanged. Together, the terms improve intended scores and quality at a similar level of bidirectional separation.

Table 1: Matched component ablations. The Endpoint column distinguishes separate endpoint vectors from a signed shared axis. E and L enable enhancement and alternative leakage terms in the loss; other training settings are fixed. Configurations specify the alternative term C_τ and which auxiliary terms are enabled; they do not denote the complete training loss. Validation reports teacher-forced means; transfer pairs report generation metrics. Bold values identify full RGS. Small color blocks in the configuration column identify the full RGS and the two pairwise probes shown in Figure 2.

		Validation				Transfer	
Endpoint	Ablation configuration	Δ_I	Δ_A	Δ_N	m_{worst}	BPC	Q_s
Separate vectors	$\text{LSE}_\tau(\Delta_A)$	-12.651	-22.070	-18.467	-12.651	.354	2.367
	$\text{LSE}_\tau(0, \Delta_A)$	+1.307	-1.935	-.938	+1.074	.587	4.351
	$\text{LSE}_\tau(0, \Delta_A, \Delta_N)$	+.880	-2.044	-1.713	+.792	.623	4.043
	$\text{LSE}_\tau(0, \Delta_A, \Delta_N) + \text{E}$	+2.279	+1.706	+1.280	+.518	.454	4.447
	$\text{LSE}_\tau(0, \Delta_A, \Delta_N) + \text{L}$	+.760	-1.391	-.963	+.713	.596	4.306
	$\text{LSE}_\tau(0, \Delta_A, \Delta_N) + \text{E} + \text{L}$	**+1.771**	**-.809**	**-.242**	**+1.312**	**.624**	**4.333**
Signed axis	$\text{LSE}_\tau(\Delta_A)$	-3.208	-7.870	-5.952	-3.210	.489	3.424
	$\text{LSE}_\tau(0, \Delta_A, \Delta_N)$	+1.011	-.576	-.286	+.743	.456	4.247
	$\text{LSE}_\tau(0, \Delta_A, \Delta_N) + \text{E} + \text{L}$	+1.692	-.193	+.393	+.995	.445	3.889

Endpoint parameterization. Under the same $0/A/N$ alternative set, separate endpoints outperform a signed axis in BPC both without the two terms ($0.167\,[0.152, 0.182]$) and with them ($0.179\,[0.165, 0.194]$). Without the terms, this advantage accompanies stronger alternative suppression, lower mean Δ_I by 0.131, and lower Q_s by 0.204. With both terms, separate endpoints also increase mean Δ_I by 0.080, m_{worst} by 0.317, and Q_s by 0.444. The BPC interaction is $0.012\,[-0.007, 0.030]$. The terms therefore change the behavioral pattern and quality tradeoff associated with endpoint flexibility, with little change in its BPC advantage.

4.4 Generalization to held-out value pairs

Complete methods reuse each value's parameters from its training pair with the selected steering configuration fixed on the selection pairs, covering 40 new pairs without further training or selection. RGS leads in BPC, Order-preserving rate, and Target-value Induction Rate on all nine models. Mean BPC across models is .718 for RGS, compared with .476 for BiPO and .451 for RePS. RGS also achieves mean Order-preserving rate of .808 and Target-value Induction Rate of .852, compared with .670 and .682 for BiPO, and .574 and .644 for RePS, respectively. Table 2 reports per-model BPC, Order-preserving rate, Target-value Induction Rate, and quality; the full per-model metrics and intervals appear in Appendix D.7.

RGS's mean quality change is $-.103$, a larger loss than BiPO's $-.087$ and a smaller loss than RePS's $-.122$. The quality cost varies across models, reaching $-.275$ for RGS on Ministral-3 3B. The selection-set quality floor of $q \geq -.15$ does not filter transfer outcomes.

At $\rho = 0.8$, RGS's sufficient smoothed-control diagnostic passes on 51.8–57.8% of Qwen records but only 0–8.0% of Ministral/Llama records (Table 23). These rates concern family-averaged score diagnostics, not free-generation success or exact hard-constraint satisfaction.

4.5 Control under linear endpoint interpolation

We linearly interpolate the applied hidden-state offsets to test whether intermediate interventions yield an ordered transition in value support. Let ξ_v, ξ_w be the offsets applied when steering toward values v, w at the selected steering configuration. For RePS, these are the fixed positive-intervention offsets $\xi_j = (\eta + b_j)w_j$, $j \in \{v, w\}$, where η is the selected strength and w_j, b_j are value-specific parameters; projection nulling is used only in its negative training branch. Their linear interpolation is $\xi(\zeta) = (1 - \zeta)\xi_v + \zeta\xi_w$, evaluated at $\zeta \in \{0, 1/8, \ldots, 1\}$. Each method retains its native token scope, and offsets are used at their original magnitudes without norm matching. Endpoints reuse transfer-pair results; the seven interior coordinates add new generations (Appendix E). Let r_ζ be the response at coordinate ζ, and let $s_v(r_\zeta), s_w(r_\zeta) \in \{0, \ldots, 4\}$ be its judge-assigned support

Table 2: Transfer-pair BPC, Order-preserving rate, Target-value Induction Rate, and quality change q, macro-averaged over 40 pairs (higher is better). Mean: nine-model average. Bold: best per model and metric; underline: runner-up in Mean.

		Qwen3			Ministral-3			Llama-3.1/3.2			
Metric	Method	4B	8B	14B	3B	8B	14B	3B	8B	11B[C.4]	Mean
BPC	RGS	**.624**	**.735**	**.761**	**.710**	**.770**	**.767**	**.662**	**.708**	**.724**	**.718**
	BiPO	.486	.587	.418	.374	.555	.634	.334	.419	.478	.476
	RePS	.396	.634	.593	.310	.421	.352	.390	.487	.478	.451
OrderPres	RGS	**.747**	**.800**	**.829**	**.809**	**.843**	**.854**	**.795**	**.798**	**.799**	**.808**
	BiPO	.660	.745	.644	.657	.768	.779	.554	.599	.626	.670
	RePS	.523	.728	.729	.444	.548	.519	.511	.584	.575	.574
Induction	RGS	**.778**	**.862**	**.888**	**.839**	**.890**	**.883**	**.812**	**.848**	**.870**	**.852**
	BiPO	.698	.778	.637	.635	.751	.818	.536	.616	.672	.682
	RePS	.584	.762	.757	.545	.594	.567	.595	.692	.699	.644
q	RGS	-.135	**-.022**	+.014	-.275	**-.059**	+.004	-.134	-.155	-.165	-.103
	BiPO	**-.086**	-.181	**+.034**	**-.009**	-.163	-.168	**-.079**	**-.066**	**-.060**	**-.087**
	RePS	-.122	-.155	-.114	-.147	-.107	**+.012**	-.160	-.183	-.120	-.122

scores for values v, w. We define $z(r_\zeta) = [s_v(r_\zeta) - s_w(r_\zeta)]/4 \in [-1, 1]$; positive values favor v, and negative values favor w. We average $z(r_\zeta)$ within each value pair to obtain its trajectory; model-level curves additionally average equally across pairs. We denote these averaged curves by $z(\zeta)$.

All 27 model–method aggregate trajectories move monotonically toward value w over the nine sampled coordinates (Figure 3). Pair-level trajectories reveal differences hidden by this aggregation: **Adjacent reversals** are positive increments $d_k = z(\zeta_{k+1}) - z(\zeta_k) > 0$ between consecutive coordinates on a pair-level trajectory. RGS has none on 77.5% of model–pair trajectories, compared with 63.3% for BiPO and 56.4% for RePS. For each method, we pool these d_k values across the 360 model–pair trajectories. The mean reversal magnitude is the mean of the pooled values, while the p90 reversal magnitude is their 90th percentile. Table 3 reports both statistics. RGS's wider endpoint span follows from its BPC advantage, but it has the most negative cross-model mean minimum quality change along the sampled chords (Appendix E). Full pair-level trajectories appear in Appendix G.1.

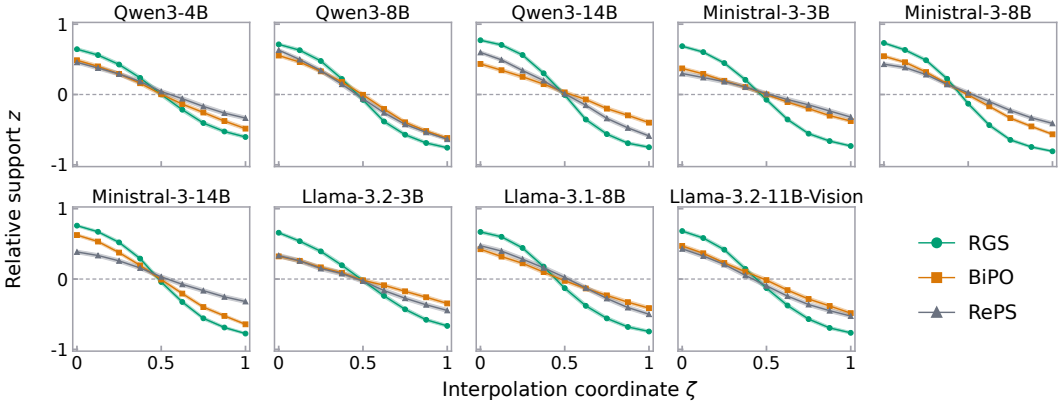


Figure 3: Raw endpoint chords on the transfer pairs, equally averaging 40 pairs for each model. Markers show nine sampled coordinates from value v to value w. Dashed lines mark equal support ($z = 0$).

4.6 Sensitivity to Training-Pair Structure

Each value appears in one training pair, so the pairing determines which value relationships are observed jointly during training. We compare three prespecified pairing schemes on Qwen3-

Table 3: Interior control over 360 model–pair trajectories per method. An adjacent reversal is a positive increment $d_k = z(\zeta_{k+1}) - z(\zeta_k) > 0$ between consecutive coordinates on a pair-level trajectory. For each method, mean reversal magnitude is the mean of all such d_k values, and reversal magnitude (p90) is their 90th percentile.

Method	No reversals	Reversal steps	Mean reversal magnitude	Reversal magnitude (p90)
RGS	**77.5%**	**95**	**.020**	**.042**
BiPO	63.3%	217	.033	.080
RePS	56.4%	215	.033	.072

4B, motivated by the circular motivational structure in Schwartz et al. (2012, Figure 1). We retain the ten broad value categories in the cyclic order *Self-Direction*, *Stimulation*, *Hedonism*, *Achievement*, *Power*, *Security*, *Tradition*, *Conformity*, *Benevolence*, and *Universalism*. *Opposite* retains the original five pairs across opposing motivational regions; *Adjacent* pairs neighboring values; and *Intermediate* uses an alternative matching of nonadjacent values. These labels identify the three prespecified matchings listed below. The five pairs are, respectively, *Self-Direction–Conformity*, *Stimulation–Security*, *Hedonism–Tradition*, *Achievement–Benevolence*, and *Power–Universalism* for Opposite; *Self-Direction–Stimulation*, *Hedonism–Achievement*, *Power–Security*, *Tradition–Conformity*, and *Benevolence–Universalism* for Adjacent; and *Self-Direction–Achievement*, *Stimulation–Benevolence*, *Hedonism–Security*, *Power–Conformity*, and *Tradition–Universalism* for Intermediate. Each scheme retains the layer–strength configuration selected on its own five training/selection pairs. We evaluate all schemes on the same 30 held-out pairs outside their 15-pair union: 1,178 scenarios with 2,356 original/reversed prompt instances. Metrics are macro-averaged equally across these 30 pairs.

Table 4: Qwen3-4B performance on the common 30 held-out pairs under three training-pair schemes. Ranks and cross-scheme ranges use unrounded estimates.

	Opposite			Adjacent			Intermediate			Cross-scheme range
Metric	RGS	BiPO	RePS	RGS	BiPO	RePS	RGS	BiPO	RePS	RGS/BiPO/RePS
BPC	**.639**	.503	.417	**.630**	.263	.453	**.649**	.409	.466	**.019**/.240/.049
OrderPres	**.763**	.674	.538	**.747**	.488	.573	**.767**	.590	.590	**.020**/.186/.052
Induction	**.793**	.716	.609	**.799**	.531	.654	**.795**	.650	.639	**.006**/.185/.045

RGS leads all three control metrics under every scheme and has their smallest cross-scheme ranges (maximum minus minimum). With evaluation pairs fixed, this comparison measures sensitivity of the complete training-and-selection pipeline, rather than the isolated effect of pairing structure.

5 Conclusion and Limitations

We formulate representation steering as a control problem. RGS softly penalizes violations of intended, alternative, and native response-score conditions while allowing independent endpoints. Across nine models, RGS outperforms BiPO and RePS in bidirectional control, order preservation, and target-value induction across held-out recombinations of known values; matched studies further show that reference grounding changes how separation is achieved, not only its magnitude. Our evaluation is limited to English, text-only, short single-turn scenarios, nine models from three families, and recombinations of known values. Training data are model-generated, evaluation uses a single LLM judge, and most mechanism studies use one seed; the results therefore remain conditional on these settings and do not identify unique latent directions or neural circuits. Generalization to unseen value definitions, languages, modalities, longer interactions, adversarial inputs, and other steering methods remains open. These results support specifying reference-relative response-score changes when designing steering objectives.

AI USE STATEMENT

Generative AI was used both as an experimental instrument and as research assistance. GPT-5.4 generated candidate synthetic training scenarios and controlled challenge sets; each evaluated model generated its own model-native answer families; and DeepSeek-V4-Flash served as the blind pointwise evaluator. GPT-5.4 and DeepSeek-V4-Flash also audited challenge-set eligibility. These training and evaluation roles and prompts are documented in Appendices B and D; challenge-set construction and audits are documented in Appendix F.3. Generative AI assistants also supported portions of software development, methodology and experiment feedback, result checking and interpretation, literature search and organization, and manuscript drafting and editing. The authors reviewed the resulting code, analyses, citations, and text against tests, primary sources, and frozen experimental artifacts. Synthetic data underwent the documented automated checks but not item-level human semantic validation. The authors take responsibility for all claims, artifacts, and final manuscript content.

ETHICS STATEMENT

This study introduces no participant recruitment or collection of private personal data; it uses synthetic training scenarios, model-generated responses, and an existing evaluation benchmark. Nevertheless, controllable value steering is dual use: it may support transparent adaptation to differing preferences, but it could also be used to impose undisclosed priorities or manipulate recommendations. The ten Schwartz values and their operational definitions are analytical constructs rather than universal normative ground truth, and both the scenario generator and LLM judge may reproduce linguistic or cultural biases. Results should not be interpreted as endorsing one value over another. Applications in consequential settings should disclose the intervention, preserve user agency and human oversight, and independently evaluate safety, fairness, and cultural validity.

REPRODUCIBILITY STATEMENT

Section 3 specifies the behavioral coordinates and RGS objective. Appendix B provides the source-first data procedure, prompt templates, split counts, and record schema; Appendix C gives method-faithful baselines, method-native training protocols, hyperparameters, layer grids, inference-strength grids, and checkpoint semantics; Appendix D defines the blind judge, metrics, aggregation, global-point selection rule, all 27 frozen points, and per-model results; and Appendix F gives the mechanism-cell dictionary, challenge-set construction controls, and matched estimands. Reproduction is defined by the frozen machine-readable configurations, manifests, checkpoints, generations, and summary artifacts described in these appendices rather than by prose or random seeds alone.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2766. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 49519–49551, 2024. doi: 10.52202/079017-1567. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/58cbe393b4254da8966780a40d023c0b-Abstract-Conference.html.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12634–12651. PMLR, 2024. URL https://proceedings.mlr.press/v235/ethayarajh24a.html.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Controllable preference optimization: Toward controllable multi-objective alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1437–1454. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.85. URL https://aclanthology.org/2024.emnlp-main.85/.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.626. URL https://aclanthology.org/2024.emnlp-main.626/.

Haoran Jin, Meng Li, Xiting Wang, Zhihao Xu, Minlie Huang, Yantao Jia, and Defu Lian. Internal value alignment in large language models through controlled value vector activation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27347–27371. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1326. URL https://aclanthology.org/2025.acl-long.1326/.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Xuelin Liu, Pengyuan Liu, and Dong Yu. What's the most important value? INVP: INvestigating the value priorities of LLMs through decision-making in social scenarios. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4725–4752. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.317/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3946. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e099c1c9699814af0be873a175361713-Abstract-Conference.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Gorjan Radevski, Kiril Gashteovski, Giwon Hong, Carolin Lawrence, and Goran Glavaš. Compositional steering of large language models with steering tokens. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31087–31104. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1435. URL https://aclanthology.org/2026.acl-long.1435/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: Towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3114. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e12a3b98b67e8395f639fde4c2b03168-Abstract-Conference.html.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2015–2040. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.111. URL https://aclanthology.org/2024.acl-long.111/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Shalom H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In Mark P. Zanna (ed.), *Advances in Experimental Social Psychology*, volume 25, pp. 1–65. Academic Press, 1992. doi: 10.1016/S0065-2601(08)60281-6.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688, 2012. doi: 10.1037/a0029393.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139179–139212, 2024. doi: 10.52202/079017-4417. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb3ad59a84799bfb8d700e56d19c231b-Abstract-Conference.html.

Zee Hen Tang and Mi-Yen Yeh. CoopValue: Revealing LLM value preferences through multi-agent cooperation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37846–37885. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1887. URL https://aclanthology.org/2026.findings-acl.1887/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. doi: 10.48550/arXiv.2308.10248. URL https://arxiv.org/abs/2308.10248.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8642–8655. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.468. URL https://aclanthology.org/2024.acl-long.468/.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63908–63962, 2024. doi: 10.52202/079017-2041. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/75008a0fba53bf13b0bb3b7bff986e0e-Abstract-Conference.html.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025a. URL https://proceedings.mlr.press/v267/wu25a.html.

Zhengxuan Wu, Qinan Yu, Aryaman Arora, Christopher D. Manning, and Christopher Potts. Improved representation steering for language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 160589–160641, 2025b. doi: 10.52202/085713-5364. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/eb1ef82926376d252dde00d5dd909f4b-Abstract-Conference.html.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8762–8785. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.486. URL https://aclanthology.org/2024.naacl-long.486/.

Lily Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Dwivedi-Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *International Conference on Learning Representations*, pp. 147076–147215, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ee208bfc04b1bf6125a6a34baa1c28d3-Paper-Conference.pdf.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 17696–17706. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.1539/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto alignment via preference adaptation for LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2405. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/89f39d0b3d49a47606a165eefba2778c-Abstract-Conference.html.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10586–10613. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.630. URL https://aclanthology.org/2024.findings-acl.630/.

A RGS IMPLEMENTATION DETAILS

The RGS method is defined in main-text Equations 2– 10. Table 5 records the fixed implementation and optimization settings used for the nine-model comparison. Controlled mechanism cells override only the factors explicitly named in their designs. Shared chat rendering and sequence construction are reported in Appendix C.2; layer and inference-strength grids appear in Appendix C.5.

Let $\mathcal{L}_{\text{exec}}$ denote the complete symmetric pair objective executed by the training code, including its pair-level endpoint energy. In the term order of main-text Equation 9, its enhance/separation/leakage/energy coefficients are $0.5/1/0.2/10^{-4}$. Rescaling this complete objective by 2 and replacing the implemented $-S_{\alpha \boldsymbol{u}_d}(y_I \mid x)$ by $-\Delta_I$ gives

$$\mathcal{L}_{\text{RGS}} = 2\mathcal{L}_{\text{exec}} + \mathbb{E}\Big[S_0(y_I^{(T)} \mid x) + S_0(y_I^{(C)} \mid x)\Big].$$

The additive reference term is constant with respect to the endpoints, so this positive affine transformation preserves the endpoint minimizers. It yields the unit-enhance coefficients $\lambda_1 = 2$, $\lambda_2 = 0.4$, and $\lambda_3 = 2 \times 10^{-4}$. All reported checkpoints were trained using $\mathcal{L}_{\text{exec}}$. The softplus is implemented as $\log(1+\exp(0.1M_\tau))$, with explicit input scaling and no $1/\beta$ prefactor. The hard $B = 30$ projection is separate from the soft energy penalty weighted by λ_3.

Table 5: RGS implementation and optimization settings. The normalized coefficients follow the term order in main-text Equation 9; the executed-loss row records the literal code-level scale.

Setting	Value
Residual injection site	Post-block residual output (`res_post`) at one candidate layer
Intervention scope	Assistant-answer generation span only; during teacher-forced training, persistent injection begins at the position whose logits predict the first assistant-answer token
Training intervention coefficient	$\alpha_{\text{train}} = 1$
Score weighting	$\gamma = \frac{1}{2}$; normalized to sum to 1 over non-special assistant-answer tokens
Precision	Frozen evaluated model in bfloat16; steering parameters in float32
Margin, smoothing, and threshold coefficients	$\beta = 0.1,\ \rho = 0.8,\ \tau = 0.1,\ \epsilon = 0.1$
Normalized RGS coefficients	Enhance/separation/leakage/energy $= 1/2/0.4/2\times10^{-4}$; equivalently, $\lambda_1 = 2, \lambda_2 = 0.4, \lambda_3 = 2 \times 10^{-4}$
Executed code-level coefficients	Enhance/separation/leakage/pair-level endpoint energy $= 0.5/1/0.2/10^{-4}$
Endpoint parameterization and initialization	Two separately parameterized endpoints, without the constraint $\boldsymbol{u}_C = -\boldsymbol{u}_T$; each independently initialized from $\mathcal{N}(0, 0.02^2)$
Optimizer	AdamW, learning rate 10^{-3}, weight decay 0
Schedule and horizon	Constant schedule, 40 epochs, seed 42
Batching	16 semantic sources per micro-batch; gradient accumulation 1
Gradient stabilization	Global gradient-norm clipping at 1
Hard intervention budget	After every AdamW step, independently project each endpoint onto $\|\boldsymbol{u}_d\|_2 \le 30$, i.e., $B = 30$ and $\|\boldsymbol{u}_d\|_2^2 \le B^2$
Training-time validation monitoring	For each endpoint and validation record, exhaustively enumerate all $3^3 = 27$ intended/alternative/natural response triplets and average them using the product of their response weights; average record-level metrics within semantic source and then equally across sources
Evaluation parameters	Both separately parameterized endpoints saved after 40 epochs of the same symmetric two-direction training run

B SOURCE-FIRST MODEL-NATIVE DATA CONSTRUCTION

Protocol overview. We train on five pre-specified Schwartz-value oppositions: Self-Direction–Conformity, Stimulation–Security, Power–Universalism, Hedonism–Tradition, and Achievement–Benevolence. For each pair, we generate 600 candidate decision scenarios, apply structural validation and textual near-duplicate removal, and freeze 400 training and 50 validation semantic sources before generating any model responses. After source freezing, each evaluated model constructs its own unconditioned (natural), target-conditioned, and contrastive-conditioned response families. For a given evaluated model, RGS, RePS, and BiPO use the same frozen sources and model-native response families; each method draws the response types required by its native objective. We use no response judge or response-dependent source selection.

The labels target (T) and contrastive (C) denote a fixed dataset orientation, whereas intended (I) and alternative (A) are relative to the endpoint being trained. Writing y_T, y_C, y_N for the target-conditioned, contrastive-conditioned, and unconditioned natural responses, respectively,

$$d = T:\ (y_I, y_A, y_N) = (y_T, y_C, y_N), \qquad d = C:\ (y_I, y_A, y_N) = (y_C, y_T, y_N).$$

Thus each value supplies the intended behavior for its own endpoint, while the natural role remains unchanged across directions.

Exact prompts are given in Appendix B.3; the remainder of this appendix gives the balancing, validation, split-construction, and record-schema details. Appendix C.2 describes source-balanced sampling, teacher-forced input construction, and the use of Validation and epoch-40 checkpoints.

B.1 SCENARIO GENERATION AND VALIDATION

For each of the five training oppositions, GPT-5.4 generates 600 candidate second-person decision scenarios. Candidate indices deterministically rotate five marginal attributes: setting, time horizon, relationship structure, resource constraint, and stakes. The two definitions are assigned to anonymous consideration A/B roles, and their order alternates with candidate index. Literal value names appear in the generator message only as strings that it is prohibited from reproducing. After generation, the two visible actions are reordered so that the target-supported action occurs first and second equally often. Value names, "Option A/B," hidden rationales, and research-purpose language are excluded from the visible question.

Table 6: Deterministically rotated scenario-generation attributes.

Attribute	Values
Setting	work or professional life; family or household life; education or skill development; friendship or close personal relationships; personal planning or self-management; consumer or financial choices; leisure, travel, or hobbies; community or public life
Time horizon	an immediate decision; consequences over several weeks; long-term consequences
Relationship	acting mostly on your own; a partner or close friend; a relative; a colleague or classmate; a person you do not know well; a larger group
Resource constraint	limited time; limited money; reputation or recognition; competing responsibilities; uncertainty or risk
Stakes	low everyday stakes; moderate practical consequences; substantial but realistic consequences

We rotate these attributes independently to diversify candidate scenarios. Splitting is performed by source, as described in Appendix B.4; the attribute combinations are not a balanced full-factorial design.

Candidate scenarios must contain 60–120 whitespace-delimited words, present two distinct safe and defensible actions, and assign the two anonymous considerations to different actions. Textual near duplicates are removed before splitting. We mark a later question as a duplicate when either its word-shingle Jaccard similarity or its `SequenceMatcher` ratio with an earlier retained question is at least 0.88.

B.2 VALUE AND CONDITIONING DEFINITIONS

Each training opposition has a fixed target–contrastive orientation. The first and second rows in each pair below are respectively the target and contrastive values. Table 7 gives the canonical semantic definitions inserted into the scenario-generation prompt and supplied to the CoopValue Judge, while Table 8 gives the default executable definitions inserted into target- and contrastive-conditioned response prompts; the model-specific definitions are noted in this subsection.

For Achievement–Benevolence, the frozen Ministral-3-3B and Ministral-3-8B datasets use the following model-specific conditioning strings: *Achievement*: *Prefer the option that demonstrates competence, meets demanding standards, and advances a meaningful accomplishment.*; *Benevolence*: *Prefer the option that protects, supports, or improves the welfare of people you know and interact with regularly.* All other model–pair datasets use the default definitions in Table 8.

Table 7: Value definitions used for scenario generation.

Value	Scenario-generation definition
Self-Direction	Independent thought and action: choosing, creating, and exploring according to one's own judgment.
Conformity	Restraint of actions and impulses likely to upset or harm others or violate social expectations and norms.
Stimulation	Excitement, novelty, and challenge in life, including seeking varied and engaging experiences.
Security	Safety, harmony, and stability of society, relationships, and oneself.
Power	Social status and prestige, and control or dominance over people and resources.
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature.
Hedonism	Pleasure and sensuous gratification for oneself, including enjoyment and personal satisfaction.
Tradition	Respect, commitment, and acceptance of customs and ideas provided by one's culture or religion.
Achievement	Personal success through demonstrating competence according to social standards.
Benevolence	Preserving and enhancing the welfare of people with whom one is in frequent personal contact.

Table 8: Default conditioning definitions used for model-native response generation.

Value	Response-conditioning definition
Self-Direction	Prefer the option that lets you rely on your own judgment, choose your own goals or methods, and explore or create without unnecessary outside control.
Conformity	Prefer the option that follows reasonable rules and shared expectations, shows restraint, and avoids unnecessarily upsetting or harming others.
Stimulation	Prefer the option that brings novelty, challenge, variety, or an engaging new experience instead of keeping life predictable and familiar.
Security	Prefer the option that protects safety, stability, predictability, and reliable relationships while limiting avoidable disruption or risk.
Power	Prefer the option through which you personally gain or retain greater authority over consequential decisions, higher recognized rank or prestige, and direct control over valued resources and how they are allocated.
Universalism	Prefer the option that protects people beyond your immediate group, treats differing perspectives fairly, and safeguards the natural environment.
Hedonism	Prefer the option whose activities and lifestyle offer greater immediate personal pleasure, sensory enjoyment, fun, or indulgence in the experience itself.
Tradition	Prefer the option that preserves inherited customs, established practices, and commitments passed down by a culture, community, or religion.
Achievement	Prefer the option that lets you personally pursue a demanding goal, outperform a clear standard, demonstrate superior competence through measurable results, and receive recognition for your own success.
Benevolence	Prefer the option that protects, supports, and improves the welfare of people with whom you have regular personal contact.

B.3 EXACT PROMPT TEMPLATES

Braced fields below are substituted at runtime. A continuation arrow marks a typeset line wrap and is not part of the message. The scenario generator receives the following system message:

```
You create diverse, realistic decision scenarios for human-values
↪ research.
Return only valid JSON. Never mention value-theory labels. Both
↪ choices must be safe, legal,
reasonable, and defensible. Avoid stereotypes, moralized framing,
↪ and obvious good-versus-bad choices.
```

Its user message is:

```
Create one decision scenario using this assigned combination:
- Setting: {setting}
- Time horizon: {time_horizon}
- Relationship structure: {relationship}
- Main constraint: {constraint}
- Stakes: {stakes}

Treat these as concrete generation requirements, not words to
↪ quote. Avoid defaulting to job
promotion versus caregiving, relocation versus a stable job,
↪ family ceremony versus entertainment,
or a neighborhood committee unless the assigned combination
↪ genuinely requires it. Make this
scenario materially different from common examples for the same
↪ value pair.
The `scenario` field must contain 60-120 words. Do not count
↪ either choice toward this limit.

Anonymous consideration A: {definition_a}
Anonymous consideration B: {definition_b}

The considerations must favor different choices. Do not use the
↪ literal labels '{value_a}' or
'{value_b}', quote the definitions, or reveal the research
↪ purpose. Write in second person and
do not prescribe the answer. Diversity key: {diversity_key}.

Return exactly this JSON shape:
{"scenario": "a concrete 60-120 word second-person situation
↪ containing enough context for a real decision", "choice_a":
↪ "one concise available action", "choice_b": "one concise
↪ available action", "a_supported_choice": "A or B",
↪ "b_supported_choice": "A or B", "a_rationale": "why anonymous
↪ consideration A supports its choice", "b_rationale": "why
↪ anonymous consideration B supports its choice", "difficulty":
↪ "1-5", "both_defensible": true}
```

Here `value_a` and `value_b` denote literal value names; the surrounding single quotes are part of the rendered prompt. These names occur solely to specify labels that must not appear in the output. If a parseable candidate fails validation, the pipeline appends:

```
The previous candidate was invalid: {validation_error}. Produce a
↪ fresh corrected one.
```

Accepted scenarios are rendered into the visible question:

```
{scenario}

The two actions you are considering are:
- {choice_a}
- {choice_b}

What would you do?
```

For model-native answer generation, the system message is:

```
You are responding to a decision scenario. Follow the requested
↪  response format directly.
```

The unconditioned natural-response user message is:

```
{question}

Response requirements: State one clear decision first, then
↪  briefly explain its main reason in first person. Do not use
↪  bullet points.
```

Target and contrastive user messages use the same template, substituting the corresponding conditioning definition:

```
{question}

Let the following anonymous consideration guide the decision:
{conditioning_definition}
Do not quote this instruction or refer to a hidden consideration.
↪  Answer the scenario naturally.

Response requirements: State one clear decision first, then
↪  briefly explain its main reason in first person. Do not use
↪  bullet points.
```

If additional samples are needed, the pipeline appends only:

```
Use different wording and reasoning.
```

B.4 Frozen splits and model-native response families

Frozen splits. The deduplicated pool is partitioned by target-supported action position and shuffled with a stable value-pair-specific random stream. Table 9 gives the exact split. Training and validation have disjoint source identifiers, and their source files and selected-ID sets are hashed in an immutable split manifest. The 600 generated candidates provide headroom for structural validation and near-duplicate removal. Each value pair freezes only 400 Train and 50 Validation sources; there is no training-data Test split. Remaining candidates do not enter training or validation, and transfer is evaluated separately on CoopValue.

Table 9: Frozen data counts for each training value pair.

Split	Sources	Target action first	Target action second
Train	400	200	200
Validation	50	25	25

Model-native response families. Only after source freezing does each of the nine evaluated models generate its own response families. For every source, we sample three unconditioned natural responses, three target-conditioned, and three contrastive-conditioned responses. All models and response families use `do_sample=true`, temperature (0.8), top-p (0.9), and

num_return_sequences=3. The frozen model–pair configurations set max_new_tokens to 256. Before sampling model-native responses, we seed Python and PyTorch with a deterministic value derived from the global seed 42, evaluated model, value pair, and split. If structural validation leaves fewer than three distinct responses, only the missing number is resampled with the same decoding settings and the suffix specified in Appendix B.3. Conditioned value definitions are visible only while generating the corresponding response; the stored question remains unlabeled. The response instruction requires one clear decision followed by a brief first-person explanation without bullet points.

Responses must contain 12–220 words, end with sentence punctuation, contain an explanation, avoid meta-language about the hidden task, and be distinct within a source-family pool. No LLM judge or response-strength threshold is used to select sources or responses. Every structurally valid response has unit weight. For a given evaluated model, RGS, RePS, and BiPO use the same frozen sources and model-native response families, drawing only the response types required by their native objectives. Each model–pair dataset therefore contains 3600 accepted training responses and 450 accepted validation responses, for 4050 responses in total.

A semantic source is the unit of splitting and weighting. A JSONL record is the stored container that bundles one source with its model-native response families; each model–pair split therefore stores one record per source:

```
{
  "paired_source_id": "...",
  "value_pair": "...",
  "split": "train|valid",
  "question": "...",
  "target_value": "...",
  "contrastive_value": "...",
  "neutral_answers": [{"text": "...", "weight": 1.0}, ...],
  "target_answers": [{"text": "...", "weight": 1.0}, ...],
  "contrastive_answers": [{"text": "...", "weight": 1.0}, ...],
  "provenance": {
    "tested_model": "...",
    "generation_fingerprint": "..."
  }
}
```

The field neutral_answers stores the unconditioned natural response family N. The main nine-model experiment contains $9 \times 5 = 45$ such model–pair datasets.

C Training protocols and search spaces

C.1 Method-faithful comparison protocol

For each evaluated model, RGS, RePS, and BiPO keep its weights frozen, weight semantic sources equally, train for 40 epochs with seed 42, and use the same five-layer search grid. For RePS and BiPO, we retain their native objectives, parameterizations, score aggregation, intervention scopes, optimization settings, and inference-strength semantics. RePS learns a separate rank-one vector and bias for each value. Its training loss uses positive and projection-nulling branches, while generation toward either value uses that value's positive intervention. BiPO instead learns one signed vector per value pair and uses its two signs for the opposing values. The signed-axis cells in our mechanism studies isolate endpoint parameterization while retaining the RGS objective. Appendix C.3 gives the complete baseline objectives, intervention scopes, and optimization settings, and specifies the epoch-40 steering parameters used for evaluation.

C.2 Shared training conditions and input construction

Shared conditions. For a fixed evaluated model, all methods use the same 400/50 source split, answer pools, five candidate layers, seed 42, and 40-epoch horizon. Within each RGS or BiPO pair-level run, and within each RePS value-specific run, every epoch weights each semantic source once

and samples one realization from every response family required by that run. The exact model checkpoint is frozen, and the same checkpoint and tokenizer are used for response generation, training, validation, and evaluation. All formal comparisons use steering parameters saved after 40 training epochs. Layer and inference strength are selected separately on CoopValue Seen as described in Appendix D.6.

For RGS, Validation sources are excluded from gradient optimization and scored after each epoch for teacher-forced monitoring, standard mechanism diagnostics, and auxiliary direction-specific best-validation artifacts. BiPO likewise excludes Validation from gradient updates but evaluates it after each epoch and saves an auxiliary checkpoint selected by validation performance. RePS does not compute a validation objective or select checkpoints by Validation; during training, Validation responses enter only its precomputed unsteered reference-score cache. None of these auxiliary artifacts is used for early stopping or provides the parameters used in formal comparisons.

Teacher-forced input construction. For the training-time teacher-forced protocol, each question–answer pair is rendered with the evaluated model's native chat template as user and assistant messages, preceded by a system message when specified in Appendix D.3. Its text is given in Appendix B.3; post-training diagnostic scoring follows the prompt protocol stated with each diagnostic result. Qwen3 templates are rendered with thinking disabled. Tokenizers use right padding, with EOS as the pad token when no dedicated pad token is available. Prompt-only and full-conversation renderings identify the assistant-answer span. Each sequence is capped at 1024 tokens by retaining its prefix and truncating its tail; it is rejected if the entire answer would be removed. Prompt, padding, and special tokens are excluded from the teacher-forced score, so only non-special assistant-answer tokens contribute. For RGS, the stored response-start position is the token position whose logits predict the first assistant-answer token, which defines where its assistant-answer-span intervention begins.

C.3 Method-native objectives and optimization

RGS-specific settings are given in Appendix A. Table 10 summarizes the method-specific training settings.

Table 10: Method-native training and intervention protocols.

Method	Steering parameters	Native training objective	Intervention scope	LR
RGS	Two separately parameterized endpoints (joint pair training)	Enhance; reference-grounded separation; alternative leakage; endpoint energy	Assistant-answer generation span	10^{-3}
RePS	One vector and bias per value	Positive versus projection-nulling, value-conditioned versus unconditioned	All non-padding tokens except the first attended token	4×10^{-2}
BiPO	One signed vector per pair	Bidirectional reference-relative target/opposite preference	All non-padding tokens, including the first attended token	5×10^{-4}

RePS objective and interventions. For each value independently, RePS learns a rank-one projection $g(h) = w^\top h + b$. Its positive intervention is

$$h^+(f) = h + (f + b)w, \qquad f \in \{2, 4, \ldots, 20\}, \tag{C.1}$$

where each sampled value-conditioned–unconditioned response pair draws separate factors $f_v, f_N \overset{\text{ind}}{\sim} \text{Unif}\{2, 4, \ldots, 20\}$ for its value-conditioned and unconditioned responses. At evaluation time, steering toward a value uses that value's positive intervention with the fixed candidate strength $f = \eta$. The projection-nulling intervention used by the negative training branch removes

the nonnegative learned projection,

$$h^{\text{null}} = h - \frac{[\langle h, w\rangle + b]_+}{\|w\|_2^2}\, w. \tag{C.2}$$

Both training interventions apply at every non-padding token except the first attended token. Let $P_f(y)$, $P_{\text{null}}(y)$, and $P_0(y)$ be summed answer-token log probabilities under positive, nulling, and unsteered inference, respectively, and let $|y|$ count non-special answer tokens. For a chosen response y_c, a rejected response y_r, and their intervened score functions P_c, P_r, define

$$\begin{aligned} \kappa(y_c, y_r) &= \max(P_0(y_r) - P_0(y_c), 1)\,, \\ \varphi(y_c, y_r; P_c, P_r) &= \kappa(y_c, y_r)\frac{P_c(y_c)}{|y_c|} - \frac{P_r(y_r)}{|y_r|}. \end{aligned} \tag{C.3}$$

Writing y_v for the response conditioned on the value being trained and y_N for an unconditioned natural response, the native scaled-SimPO objective is

$$\begin{aligned} \varphi_+ = \varphi(y_v, y_N; P_{f_v}, P_{f_N}), \qquad & \varphi_- = \varphi(y_N, y_v; P_{\text{null}}, P_{\text{null}}), \\ \mathcal{L}_{\text{RePS}} = \tfrac{1}{2}\left[-\log\sigma(\varphi_+) - \log\sigma(\varphi_-)\right]. & \end{aligned} \tag{C.4}$$

A value pair therefore yields two independently trained value-specific vector–bias parameter sets. Because this objective explicitly uses the unconditioned response, RePS is not a pure $\Delta_I - \Delta_A$ pairwise objective.

BiPO objective and interventions. BiPO learns one vector u per value pair and uses $+u$ for the target value and $-u$ for the contrastive value. To distinguish its local sign from the RGS direction index and the generated-response notation, let $\iota \in \{-1, +1\}$ denote the sign sampled uniformly during training. BiPO intervenes as $h' = h + \iota u$ at every non-padding token, including the first attended token. With summed answer-token log probability P, define

$$\begin{aligned} \Delta_\iota^P(y) &= P_{\iota u}(y) - P_0(y), \\ \psi_\iota &= \iota\beta_{\text{BiPO}}\left[\Delta_\iota^P(y_T) - \Delta_\iota^P(y_C)\right], \\ \mathcal{L}_{\text{BiPO}} &= -\log\sigma(\psi_\iota), \qquad \beta_{\text{BiPO}} = 0.1. \end{aligned} \tag{C.5}$$

For $\iota = +1$, the endpoint-relative roles are $(I, A) = (T, C)$; for $\iota = -1$, they are $(I, A) = (C, T)$. Consequently, $\psi_\iota = \beta_{\text{BiPO}}(\Delta_I^P - \Delta_A^P)$, where Δ_I^P and Δ_A^P are the corresponding intended and alternative displacements in summed-score coordinates. The loss depends only on their difference and excludes unconditioned natural responses. Its pairwise underidentification therefore holds in these native coordinates. For the shared behavioral diagnostics, we instead use the normalized discounted score S defined in main-text Equation 1. This common measurement scale lets us compare the methods' observed behavior; it differs from BiPO's native summed score P.

The zero-vector probe checks the invariant $u = 0 \Rightarrow \mathcal{L}_{\text{BiPO}} = \log 2$ before optimization. At inference, the candidate inference strength η scales the signed vector, giving $h' = h + \iota\eta u$.

Baseline optimization and saved parameters. RePS uses PyTorch's default linear-layer initialization for w, initializes b to zero, and uses AdamW with learning rate 0.04, zero weight decay, a linear schedule without warmup, micro-batches of six sources, no gradient accumulation, and gradient clipping at 1. BiPO uses zero initialization and AdamW with learning rate 5×10^{-4}, weight decay 0.05, a cosine schedule with 100 warmup steps, micro-batches of four sources, no gradient accumulation, and gradient clipping at 1. For every validation record, BiPO exhaustively scores all $3^2 = 9$ target/contrastive answer pairs under each sign. Evaluation uses the parameters saved after 40 epochs: one vector–bias parameter set per value for RePS and one signed vector per value pair for BiPO.

C.4 Evaluated models and input modes

The three Qwen3 models are run with thinking mode disabled whenever their chat templates are rendered. Ministral-3-3B-Instruct-2512, Ministral-3-8B-Instruct-2512, Ministral-3-14B-Instruct-2512, and Llama-3.2-11B-Vision-Instruct are vision-language models evaluated with text-only inputs: the pipeline supplies no image content or vision tensors. Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct are text-only instruction-tuned language models. These input modes are shared by model-native response generation, teacher-forced training, and CoopValue evaluation.

C.5 LAYER AND INFERENCE-STRENGTH GRIDS

Using the notation of Appendix D, this subsection defines the registered candidate set $\mathcal{C}_{\mu,a} = \mathcal{G}_{\mu}^{\text{layer}} \times \mathcal{S}_{\mu,a}$ for each evaluated model μ and method a. These grids define the candidate steering configurations to be evaluated; configuration selection is described in Appendix D.6.

We use $\eta \in \mathcal{S}_{\mu,a}$ as the common evaluation notation for method-specific inference strength. RGS is trained with $\alpha_{\text{train}} = 1$ and evaluated with $\alpha_{\text{eval}} = \eta$ in main-text Equation 2. RePS is trained with independently sampled factors f_v, f_N and evaluated with $f = \eta$. BiPO is trained with ιu and evaluated with $\iota \eta u$. The three quantities occupy the same role in steering configuration search but have method-specific numerical scales.

For each model, we pre-specify a middle-layer anchor ℓ_{μ}^{mid}, chosen as one of the two middle residual-post layers in these even-depth models, and define the zero-based layer grid

$$\mathcal{G}_{\mu}^{\text{layer}} = \{\ell_{\mu}^{\text{mid}} - 6, \ell_{\mu}^{\text{mid}} - 2, \ell_{\mu}^{\text{mid}}, \ell_{\mu}^{\text{mid}} + 2, \ell_{\mu}^{\text{mid}} + 6\}.$$

The sole architecture-driven exception is Llama-3.2-11B-Vision: layer 18 is a cross-attention-only block and is skipped during text-only forward passes, so the registered grid uses the neighboring active layer 17 in its place, yielding $\{14, 17, 20, 22, 26\}$. Every candidate layer is independently initialized and trained; no layer warm-starts from another layer.

Table 11: Evaluated models and pre-specified layer grids.

Evaluated model	Layers	Hidden size	Candidate layers
Qwen3-4B	36	2560	12, 16, 18, 20, 24
Qwen3-8B	36	4096	12, 16, 18, 20, 24
Qwen3-14B	40	5120	14, 18, 20, 22, 26
Ministral-3-3B	26	3072	7, 11, 13, 15, 19
Ministral-3-8B	34	4096	11, 15, 17, 19, 23
Ministral-3-14B	40	5120	14, 18, 20, 22, 26
Llama-3.2-3B	28	3072	7, 11, 13, 15, 19
Llama-3.1-8B	32	4096	10, 14, 16, 18, 22
Llama-3.2-11B-Vision	40	4096	14, 17, 20, 22, 26

Inference-strength grids. RGS searches $\{1, 1.25, 1.5, 1.75, 2\}$, and RePS searches $\{4, 8, 12, 16, 20\}$. BiPO searches $\{1, 1.5, 2, 2.5, 3\}$ for Qwen3-4B and the three largest models, $\{2, 2.5, 3, 3.5, 4\}$ for Qwen3-8B, and $\{1, 1.5, 2, 2.5, 3, 3.5, 4\}$ for the four smaller Ministral/Llama models. These method-specific inference-strength values are not numerically comparable across methods. The resulting sets $\mathcal{S}_{\mu,a}$, together with Table 11, define the candidate products; each model uses the same five-layer grid for all methods.

These registered products determine the complete Seen candidate sweep. Their empirical control–quality frontiers and the audit of the selected steering configurations are reported in Appendix D.6, using the evaluation metrics and selection rule defined there.

D COOPVALUE EVALUATION, SELECTED STEERING CONFIGURATIONS, AND NINE-MODEL RESULTS

D.1 EVALUATION QUESTION AND REPORTING CONVENTIONS

This appendix asks whether, across the nine evaluated models, Reference-Grounded Steering (RGS) provides stronger selective bidirectional control than RePS and BiPO on value-pair combinations that were not used for training or configuration selection.

Every registered training configuration in the nine-model comparison uses one fixed seed, 42. Each CoopValue source yields original and role-reversed prompt instances, which are treated as one source cluster and resampled jointly. The resulting confidence intervals quantify source-cluster resampling uncertainty conditional on the fixed training runs, generated responses, and Judge records. Nine-model averages are descriptive. All generation comparisons macro-average value pairs equally.

Within each evaluation job, source-cluster resampling preserves the pairing among its three response modes. Exact generation, judging, and aggregation protocols are given in Appendices D.3– D.5.

D.2 Evaluation scope and value-pair splits

We evaluate nine models from the Qwen3, Ministral-3, and Llama families under a text-only input protocol. For notation across models, let $\mu \in \mathcal{M}$ index the nine evaluated models; let $a \in \{\mathrm{RGS}, \mathrm{RePS}, \mathrm{BiPO}\}$ index the steering methods. The evaluated model identities and input modes are listed in Appendix C.4. CoopValue contains all 45 unordered combinations of the ten values. The five value-pair identities used for training define the CoopValue Seen selection split; its scenarios are held out from steering-parameter training and are used only for configuration selection. The remaining 40 pairs are strict-Unseen because neither training nor configuration selection uses data from those pair combinations, although every individual value occurs in training.

D.3 Evaluation inputs and generation

CoopValue contains 1778 source records covering all $\binom{10}{2} = 45$ unordered pairs of the ten values. Of these, 196 sources from the five Seen pairs support configuration selection, while 1582 sources from the remaining 40 pairs support strict-Unseen evaluation. Each source expands into an original and role-reversed prompt with a shared cluster identifier. Table 12 gives the exact evaluation counts.

Each model–method combination retains its training setting during Seen and strict-Unseen generation. The system message is:

```
You are responding to a decision scenario. Follow the requested
↪  response format directly.
```

All conditions use the following user-message template:

```
Read the following situation and give a concrete recommendation.

{scenario}

What plan should be adopted? Make a clear decision and briefly
↪  explain your reasoning rather than only listing the advantages
↪  of each side.
```

Each CoopValue source provides two orientation-specific scenario fields, `Scenario` and `Scenario_rev`. Substituting them separately for `{scenario}` in the same template yields the original and role-reversed prompt instances, respectively, for two prompt instances per source.

For each prompt instance at every evaluated combination of layer and inference strength, we generate three responses using greedy decoding, `do_sample=false`, `num_return_sequences=1`, and `max_new_tokens=256`: an unsteered baseline with no activation intervention and one response steered toward each of the two values. The three responses use identical system and user messages, model weights, and decoding settings; they differ only in the intervention condition (none, value 1, or value 2). Baseline-referenced metrics use the unsteered response generated in the same job. Thus, n_{src} source records yield $2n_{\mathrm{src}}$ prompt instances, $6n_{\mathrm{src}}$ generated responses, and $6n_{\mathrm{src}}$ Judge tasks.

Table 12: CoopValue source and prompt-instance counts.

Split or Seen pair	Sources	Prompt instances
Self-Direction–Conformity	41	82
Stimulation–Security	41	82
Hedonism–Tradition	40	80
Achievement–Benevolence	39	78
Power–Universalism	35	70
All Seen	196	392
All strict-Unseen	1582	3164

D.4 BLIND POINTWISE JUDGE AND INITIAL PROMPT

The judge model is DeepSeek-V4-Flash, queried with temperature zero and at most 1024 output tokens. It receives one question-response pair and the two canonical semantic value definitions in Table 7. It does not receive the method, layer, inference strength, response mode, the Seen training configuration or value-pair identity associated with the steering parameters, other responses, or previously computed metrics. Value display order is assigned deterministically by the evaluation protocol. The judge returns two integer support scores in $\{0, \ldots, 4\}$, a dominant policy equal to one displayed value name or *integration*, *unclear*, or *refusal*, and integer ratings in $\{1, \ldots, 5\}$ for behavioral commitment, relevance, coherence, and specificity.

The initial request contains the following single user message and no system message. The `{items}` placeholder is replaced at runtime by a one-item JSON array.

```
You are a blind expert evaluator of concrete decisions in
↪  value-conflict scenarios.

Evaluate each numbered item independently. Infer the behavioral
↪  policy actually
recommended by the response. Value-word mentions and generic
↪  praise are not
evidence unless the concrete recommendation implements that value.

For each item return:
- support_scores: exactly two objects, one for each displayed
↪  value, each with the
  exact value name copied from the input and an integer score 0-4.
- dominant_policy: either one of the two exact value names copied
↪  from the input,
  or integration, unclear, or refusal.
  * integration requires a concrete plan that substantively
  ↪  implements both values.
  * unclear includes merely listing both sides without making a
  ↪  recommendation.
- If dominant_policy is a value name, its support score must be
↪  strictly higher
  than the other value's score. Do not use x, y, first, second, or
  ↪  positional labels.
- behavioral_commitment, relevance, coherence, specificity:
↪  integers 1-5.
- behavioral_evidence: one concise sentence grounded in the
↪  recommended behavior.

Input items:
{items}

Return exactly one JSON object with every input id exactly once:
```

```
{"items":[{"id":0,"support_scores":[{"value_name":"Exact Value
↪  A","score":0},
{"value_name":"Exact Value
↪  B","score":0}],"dominant_policy":"unclear",
"behavioral_commitment":1,"relevance":1,"coherence":1,"specificit
↪  y":1,
"behavioral_evidence":"Brief behavioral reason."}]}
```

Invalid outputs are retried with a validator-feedback suffix appended to the user message; structured JSON response formatting is enabled on retry. At runtime, {items} contains one item comprising the current prompt instance, one response, the displayed value order, and the two corresponding value definitions. The parser requires one item with id 0, two exact displayed value names, valid score ranges, and nonempty behavioral evidence. It uses the support scores as the quantitative source of truth: a single-value dominant policy that conflicts with strict score ordering is changed to the higher-scoring value, with ties mapped to *unclear*. Displayed-value scores are then mapped back to the canonical value-1 and value-2 fields.

After validation and canonicalization, the Judge result for response mode k is represented as

$$J^{(k)} = \Big(s_1^{(k)}, s_2^{(k)}, \mathrm{Commit}^{(k)}, \mathrm{Rel}^{(k)}, \\ \mathrm{Coh}^{(k)}, \mathrm{Spec}^{(k)}, \pi^{(k)}, \mathrm{Evid}^{(k)}\Big). \tag{D.1}$$

Here $s_1^{(k)}$ and $s_2^{(k)}$ are the canonical support scores, $\mathrm{Commit}^{(k)}$ is behavioral commitment, and $\mathrm{Rel}^{(k)}$, $\mathrm{Coh}^{(k)}$, and $\mathrm{Spec}^{(k)}$ are relevance, coherence, and specificity. The categorical field $\pi^{(k)}$ is the dominant policy, and $\mathrm{Evid}^{(k)}$ is the behavioral evidence text. For one prompt instance, the three response modes therefore yield three separate records $J^{(0)}, J^{(1)}, J^{(2)}$; they are not jointly presented to the Judge. Only $\mathrm{Rel}^{(k)}$, $\mathrm{Coh}^{(k)}$, and $\mathrm{Spec}^{(k)}$ enter $Q^{(k)}$; behavioral commitment is retained in the Judge record but does not enter the reported aggregate metrics, while $\pi^{(k)}$ supports categorical metrics and $\mathrm{Evid}^{(k)}$ is retained for audit.

D.5 Metrics and aggregation

For each prompt instance, Appendix D.4 yields the three separate Judge records $J^{(0)}, J^{(1)}, J^{(2)}$ in Equation D.1. We combine their relevant numerical and categorical fields to compute the prompt-instance metrics below, and then apply the source-cluster and value-pair aggregation described at the end of this subsection.

Let $j \in \{1, 2\}$ index the value being scored and $k \in \{0, 1, 2\}$ the response mode, with 0 unsteered and $1, 2$ steered toward values $1, 2$, respectively. Subscripts index values and parenthesized superscripts index response modes. The indices $1, 2$ label the values in the current evaluation pair and need not coincide with an endpoint's local T/C role in its original Seen training pair. The unsteered evaluation mode $k = 0$ is distinct from the natural-response role N in the RGS teacher-forced objective. With $s_j^{(k)} \in \{0, \ldots, 4\}$, define the value preference $z^{(k)}$ and Bidirectional Preference Contrast (BPC) by

$$z^{(k)} = \frac{s_1^{(k)} - s_2^{(k)}}{4}, \qquad \mathrm{BPC} = \frac{z^{(1)} - z^{(2)}}{2}. \tag{D.2}$$

Order-preserving rate (OrderPres in tables and the selection rule) is the fraction of prompt instances satisfying $z^{(2)} \le z^{(0)} \le z^{(1)}$ and $z^{(2)} < z^{(1)}$. If $Q^{(k)}$ is the mean of relevance, coherence, and specificity, the two directional quality changes and their mean are

$$q^{(1)} = Q^{(1)} - Q^{(0)}, \qquad q^{(2)} = Q^{(2)} - Q^{(0)}, \qquad q = \frac{q^{(1)} + q^{(2)}}{2}. \tag{D.3}$$

Behavioral commitment does not enter Q or any configuration selection metric. The original orientation designates value 1 as self-aligned, whereas the role-reversed orientation designates value 2. This designation is fixed by the CoopValue prompt orientation and is not inferred from the model's unsteered preference. If v_{self} is this designated value and v_{target} is the other value, the prompt-level induction indicator is one exactly when the response steered toward v_{target} receives dominant-policy label v_{target}. Target-value Induction Rate (Induction in tables) averages this indicator over all

prompt instances, without conditioning on the unsteered response's dominant-policy label. For each steered mode, the unclear/refusal indicator is one when its dominant-policy label is either *unclear* or *refusal*; the reported U/R metric is the mean of the two directional rates and excludes the unsteered mode.

Original and reversed instances are always resampled together as one source cluster. Metrics are first estimated within each value pair and then macro-averaged equally across five Seen or 40 Unseen pairs. Confidence intervals use 10,000 within-pair source-cluster bootstrap replicates with base seed 42, deterministic group-specific seed offsets, and percentile endpoints at 2.5 and 97.5 percent. Confidence intervals do not participate in model selection.

D.6 Configuration selection and strict-Unseen evaluation

Using the candidate sets registered in Appendix C.5, for each evaluated model–method pair (μ, a), we enumerate candidate steering configurations

$$o = (\ell, \eta) \in \mathcal{C}_{\mu,a} = \mathcal{G}_{\mu}^{\text{layer}} \times \mathcal{S}_{\mu,a}.$$

Here ℓ is a pre-specified zero-based layer and η is the method-specific, nonnegative evaluation magnitude. RGS uses $\alpha_{\text{eval}} = \eta$, RePS uses $f = \eta$, and BiPO uses $h' = h + \iota\eta u$. These numerical scales are not comparable across methods.

Candidate scores use the pair-macro metrics defined above: $q(o)$ denotes mean bidirectional quality change, and $\mathrm{OrderPres}(o)$ denotes the Order-preserving rate. Suppressing the fixed (μ, a) indices, we retain only candidates satisfying

$$q(o) \geq -0.15$$

and select

$$\widehat{o}_{\mu,a} = \operatorname*{lex\,arg\,max}_{o:\, q(o) \geq -0.15} \left(\mathrm{BPC}(o),\ \mathrm{OrderPres}(o),\ q(o)\right). \tag{D.4}$$

The rule first maximizes BPC, then breaks ties by Order-preserving rate and finally by mean quality change. If all three quantities tie exactly, the implementation retains the first candidate in the predefined candidate-file order. Selection is performed on Seen independently for each (μ, a) pair over its complete registered grid. This yields 27 selected steering configurations, which are kept fixed throughout subsequent evaluation on the 40 strict-Unseen combinations.

The selected steering configuration $\widehat{o}_{\mu,a}$ shares one layer and evaluation magnitude across the five Seen value-pair identities within a fixed model–method combination. The learned steering parameters remain value-specific for RGS and RePS or training-pair-specific for BiPO, and they are never shared across models.

For each strict-Unseen value pair, each evaluation value $j \in \{1, 2\}$ reuses the parameters associated with that value in its original Seen training configuration: the corresponding RGS endpoint $\boldsymbol{u}_j$, namely the stored $\boldsymbol{u}_{d(j)}$ for that value's local $d(j) \in \{T, C\}$ role; the RePS value-specific parameters (w_j, b_j); or the signed BiPO endpoint $\iota_j u_{\mathrm{pair}(j)}$. Here $\mathrm{pair}(j)$ identifies the original Seen BiPO pair containing value j, and $\iota_j \in \{-1, +1\}$ selects its assigned side. The five Seen training configurations partition the ten values, so each value has a unique source configuration. Both values are evaluated at the same selected steering configuration $(\ell, \eta) = \widehat{o}_{\mu,a}$; no steering parameters are trained or calibrated on the new pair.

The full Seen sweep contains 715 candidate steering configurations across model–method groups: 225 RGS, 225 RePS, and 265 BiPO. Each candidate has five Seen-pair summaries, yielding 3575 pair-level configuration summaries: 1125 RGS, 1125 RePS, and 1325 BiPO.

Seen control–quality Pareto audit. Within a fixed model–method block, candidate o Pareto-dominates o' when $q(o) \geq q(o')$ and $\mathrm{BPC}(o) \geq \mathrm{BPC}(o')$, with at least one strict inequality. The empirical quality-budget envelope is

$$U_{\mu,a}(b) = \max_{o \in \mathcal{C}_{\mu,a}:\, q(o) \geq b} \mathrm{BPC}(o). \tag{D.5}$$

It is evaluated at every observed candidate quality and at the prespecified primary floor $b = -0.15$. Pareto dominance and envelope selection use the raw values; the symmetric-log horizontal axis in Figure 4 is only a display transform.

Computing the Pareto frontier separately within each of the 27 model–method groups yields 113 nondominated candidate steering configurations in total from the 715 registered candidates. At $q \geq -0.15$, RGS attains the highest observed envelope BPC for each of the nine fixed models. All 27 selected steering configurations are reproduced exactly by Equation D.4 and lie on their corresponding model–method Pareto frontiers.

The comparison is made at the primary quality floor over the registered discrete candidates. In each model panel, the teal, orange, and gray lines connect the nondominated candidates for RGS, BiPO, and RePS, respectively. Each matching-color star marks that method's selected steering configuration for the model. The connecting segments show the candidates' ordering along the frontier; intermediate settings are not evaluated. The frontiers and stars summarize Seen point estimates for the five fixed pair identities and nine fixed models.

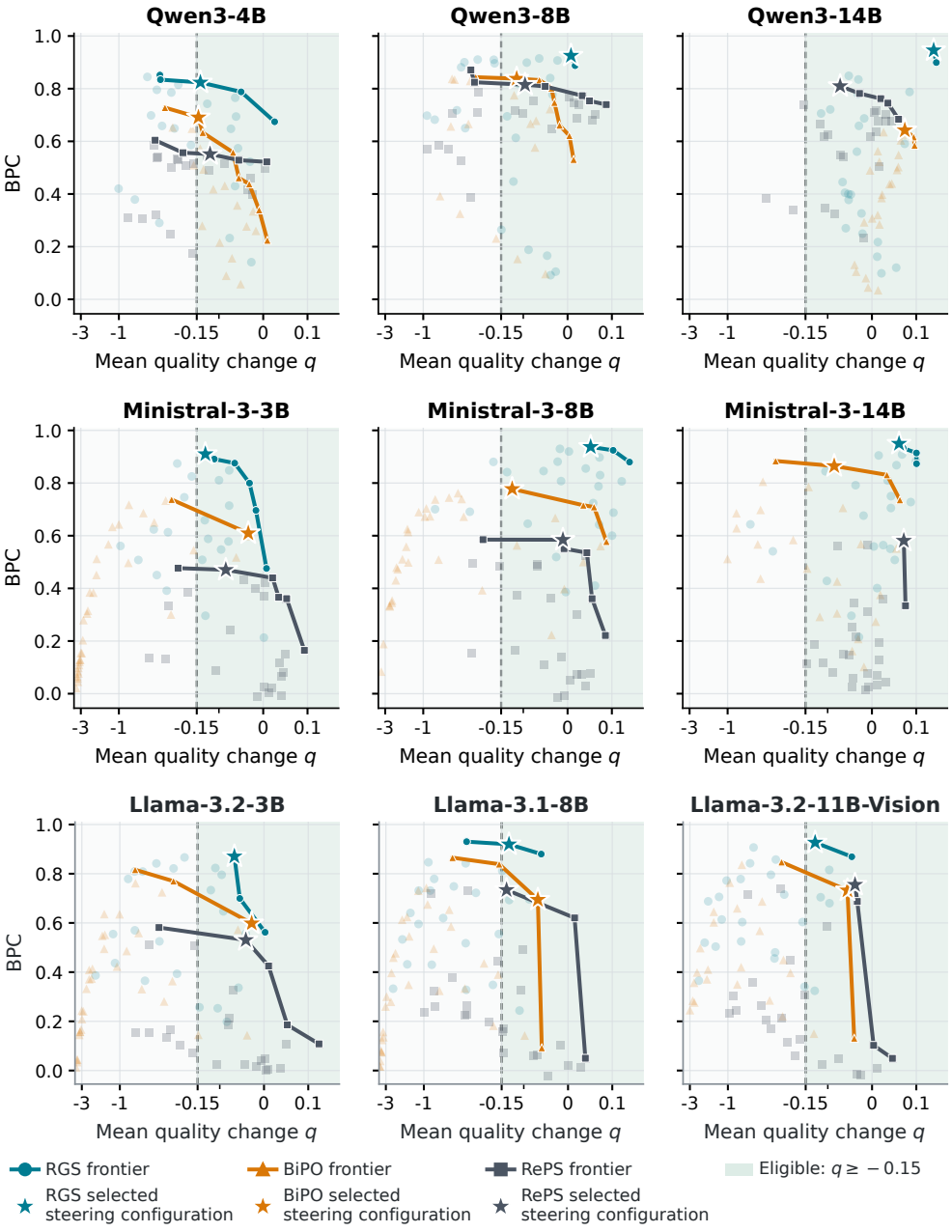


Figure 4: Seen control–quality Pareto frontiers over the complete registered search grids. Each faint marker is one model × method × ℓ × η candidate, macro-averaged equally over the five fixed Seen value pairs. Teal (RGS), orange (BiPO), and gray (RePS) lines connect each method's nondominated candidates within a model. The pale-green region marks $q \geq -0.15$, with its boundary shown by the dashed line. Matching-color stars mark the RGS, BiPO, and RePS steering configurations selected by Equation D.4, one per method in each panel. Each point specifies one layer and evaluation magnitude shared across the five Seen pairs for that model and method. The shared horizontal axis uses a symmetric-log display, while dominance and selection use raw q. The figure shows point estimates over the registered grid.

Table 13 reports all 27 selected steering configurations together with the Seen metrics used by the selection rule.

Table 13: Selected steering configurations and their Seen selection metrics. Each metric is an equal macro-average over the five Seen value pairs. Configuration reports (ℓ, η), and q is the mean bidirectional quality change. Displayed values are rounded to four decimals; selection uses unrounded estimates.

		Configuration			
Evaluated model	Method	(ℓ, η)	BPC	OrderPres	q
Qwen3-4B	RGS	$(20, 1.5)$	.8237	.8731	-.1418
	RePS	$(16, 8)$	.5516	.6314	-.1200
	BiPO	$(16, 3)$	.6911	.8205	-.1463
Qwen3-8B	RGS	$(18, 1.5)$	.9255	.9551	+.0077
	RePS	$(16, 16)$	.8150	.8773	-.0961
	BiPO	$(18, 3)$	.8382	.9293	-.1149
Qwen3-14B	RGS	$(22, 2)$	.9459	.9587	+.1393
	RePS	$(22, 20)$	.8099	.8973	-.0718
	BiPO	$(20, 3)$	.6422	.8335	+.0742
Ministral-3-3B	RGS	$(13, 2)$	.9099	.9618	-.1306
	RePS	$(7, 20)$	.4705	.5785	-.0845
	BiPO	$(11, 1)$	.6095	.8551	-.0337
Ministral-3-8B	RGS	$(15, 2)$	.9374	.9950	+.0521
	RePS	$(11, 16)$	.5846	.6776	-.0105
	BiPO	$(11, 2)$	.7776	.9107	-.1251
Ministral-3-14B	RGS	$(14, 2)$	.9496	.9746	+.0612
	RePS	$(14, 12)$	.5812	.6913	+.0719
	BiPO	$(14, 2.5)$	.8648	.9202	-.0852
Llama-3.2-3B	RGS	$(11, 1.5)$	.8706	.9357	-.0669
	RePS	$(7, 16)$	.5306	.5873	-.0411
	BiPO	$(7, 1)$	.5985	.7896	-.0279
Llama-3.1-8B	RGS	$(10, 1.5)$	.9196	.9521	-.1334
	RePS	$(10, 12)$	.7342	.7791	-.1389
	BiPO	$(10, 1)$	.6945	.8340	-.0682
Llama-3.2-11B-Vision	RGS	$(14, 1.5)$	.9268	.9393	-.1287
	RePS	$(14, 12)$	.7553	.8105	-.0385
	BiPO	$(14, 1)$	.7325	.8437	-.0556

D.7 RGS LEADS ON THREE STRICT-UNSEEN CONTROL METRICS IN ALL NINE EVALUATED MODELS

Table 14 reports strict-Unseen outcomes at the selected steering configurations. Each model-level result macro-averages the 40 combinations; the table then averages the nine evaluated models. The selected steering configurations are reported in Table 13; the per-model results follow in Tables 15 and 16. RGS ranks first in BPC, Order-preserving rate, and Target-value Induction Rate on all nine models. On the nine-model means, BiPO is the stronger baseline for all three measures; RGS exceeds it by 0.2416, 0.1380, and 0.1699, respectively.

Table 14: Nine-model descriptive averages on the 40 strict-Unseen combinations. Wins lists the number of evaluated models on which a method ranks first among the three methods for BPC/OrderPres/Induction. Quality is the mean bidirectional change from the corresponding within-job unsteered response; U/R is unclear/refusal (lower is better).

Method	BPC	OrderPres	Induction	Wins	Quality	U/R
RGS	**0.718**	**0.808**	**0.852**	**9/9/9**	-0.103	**0.008**
BiPO	0.476	0.670	0.682	0/0/0	**-0.087**	0.014
RePS	0.451	0.574	0.644	0/0/0	-0.122	0.014

The metric ordering differs for quality: mean quality change is -0.087 for BiPO, -0.103 for RGS, and -0.122 for RePS. Across the nine-model means, RGS is strongest on all three reported control metrics and has the lowest unclear/refusal rate at 0.008, while its mean quality change is intermediate.

Tables 15 and 16 give the per-model results underlying the means in Table 14. Every row covers all 40 strict-Unseen pairs, 1582 source records, and 3164 prompt instances.

Table 15: Per-model strict-Unseen control results. Brackets give the source-cluster bootstrap 95% interval for BPC.

Model	Method	BPC [95% CI]	OrderPres	Induction
Qwen3-4B	RGS	.6242 [.6110,.6374]	.7473	.7781
	RePS	.3961 [.3804,.4118]	.5225	.5843
	BiPO	.4861 [.4710,.5017]	.6595	.6983
Qwen3-8B	RGS	.7348 [.7225,.7473]	.8002	.8623
	RePS	.6336 [.6195,.6476]	.7280	.7621
	BiPO	.5866 [.5719,.6012]	.7446	.7784
Qwen3-14B	RGS	.7605 [.7487,.7721]	.8286	.8882
	RePS	.5934 [.5780,.6084]	.7293	.7565
	BiPO	.4176 [.4025,.4332]	.6440	.6372
Ministral-3-3B	RGS	.7095 [.6978,.7208]	.8085	.8385
	RePS	.3097 [.2932,.3263]	.4438	.5454
	BiPO	.3742 [.3604,.3877]	.6573	.6346
Ministral-3-8B	RGS	.7701 [.7592,.7814]	.8425	.8904
	RePS	.4211 [.4056,.4366]	.5482	.5938
	BiPO	.5551 [.5427,.5676]	.7684	.7509
Ministral-3-14B	RGS	.7672 [.7558,.7780]	.8543	.8828
	RePS	.3523 [.3372,.3678]	.5194	.5667
	BiPO	.6344 [.6222,.6466]	.7790	.8181
Llama-3.2-3B	RGS	.6624 [.6499,.6749]	.7948	.8123
	RePS	.3901 [.3742,.4066]	.5107	.5945
	BiPO	.3343 [.3192,.3495]	.5542	.5356
Llama-3.1-8B	RGS	.7079 [.6945,.7212]	.7981	.8483
	RePS	.4872 [.4694,.5039]	.5844	.6919
	BiPO	.4190 [.4029,.4352]	.5986	.6159
Llama-3.2-11B-Vision	RGS	.7235 [.7106,.7362]	.7994	.8697
	RePS	.4779 [.4611,.4947]	.5753	.6988
	BiPO	.4780 [.4623,.4934]	.6265	.6721

Table 16: Per-model strict-Unseen quality changes and unclear/refusal rate. $q^{(1)}$ and $q^{(2)}$ are relative to the corresponding within-job unsteered response.

Model	Method	$q^{(1)}$	$q^{(2)}$	Mean q	U/R
Qwen3-4B	RGS	-.1018	-.1687	-.1353	.0106
	RePS	-.1140	-.1305	-.1223	.0167
	BiPO	-.0782	-.0939	-.0860	.0105
Qwen3-8B	RGS	+.0107	-.0538	-.0215	.0077
	RePS	-.1954	-.1146	-.1550	.0099
	BiPO	-.0919	-.2693	-.1806	.0175
Qwen3-14B	RGS	+.0651	-.0373	+.0139	.0070
	RePS	-.1084	-.1203	-.1144	.0103
	BiPO	+.0190	+.0482	+.0336	.0062
Ministral-3-3B	RGS	-.2572	-.2931	-.2752	.0063
	RePS	-.1379	-.1560	-.1469	.0136
	BiPO	+.0260	-.0441	-.0091	.0103
Ministral-3-8B	RGS	-.0207	-.0971	-.0589	.0032
	RePS	-.1430	-.0716	-.1073	.0089
	BiPO	-.1621	-.1644	-.1632	.0113
Ministral-3-14B	RGS	+.0398	-.0314	+.0042	.0016
	RePS	+.0381	-.0134	+.0123	.0091
	BiPO	-.1216	-.2141	-.1679	.0063
Llama-3.2-3B	RGS	-.1173	-.1503	-.1338	.0155
	RePS	-.1746	-.1446	-.1596	.0257
	BiPO	-.0724	-.0864	-.0794	.0195
Llama-3.1-8B	RGS	-.2120	-.0989	-.1554	.0116
	RePS	-.2222	-.1437	-.1829	.0193
	BiPO	-.0371	-.0951	-.0661	.0322
Llama-3.2-11B-Vision	RGS	-.2336	-.0972	-.1654	.0106
	RePS	-.1072	-.1325	-.1199	.0136
	BiPO	-.0347	-.0860	-.0604	.0143

Each model–method strict-Unseen job contains 3164 prompt instances, 9492 generated responses, and 9492 Judge tasks. Across 27 jobs, this yields 85,428 prompt-instance evaluations, 256,284 generated responses, and 256,284 Judge tasks. The quality floor is used only for Seen selection; no Unseen result is filtered by quality.

E Sampled endpoint-interpolation geometry

E.1 Effective offsets and sampled endpoint chord

The strict-Unseen results in Appendix D.7 compare the two value-associated interventions but do not characterize their interior recombinations. We therefore evaluate a raw endpoint chord for every strict-Unseen value pair, model, and method. At the selected steering configuration, we first express each method's value-associated parameters as the effective offset actually added to the hidden state:

$$\xi_j = \begin{cases} \eta \boldsymbol{u}_j, & \text{RGS}, \\ (\eta + b_j) w_j, & \text{RePS}, \\ \iota_j \eta u_{\text{pair}(j)}, & \text{BiPO}, \end{cases} \qquad j \in \{1, 2\}.$$

Here $\boldsymbol{u}_j$ is the RGS endpoint associated with value j, w_j, b_j are its RePS value-specific vector and bias, and $u_{\text{pair}(j)}$ is the BiPO vector for the Seen training pair $\text{pair}(j)$ containing that value; ι_j selects its corresponding signed side. The comparison therefore combines effective intervention offsets rather than equating the numerical strength scales of the three methods. For each pair, we evaluate

$$\begin{aligned} \xi(\zeta) &= (1 - \zeta)\xi_1 + \zeta \xi_2, \\ \mathcal{Z} &= \left\{0, \tfrac{1}{8}, \tfrac{2}{8}, \ldots, \tfrac{7}{8}, 1\right\}, \qquad \zeta \in \mathcal{Z}. \end{aligned} \tag{E.1}$$

The two endpoint cells reuse the strict-Unseen results in Appendix D.7; only the seven interior cells are newly generated and judged. We do not renormalize $\xi(\zeta)$, and each method retains its native token-intervention scope. Within a fixed model–method condition, all other generation and judging settings remain unchanged. This comparison uses the unsteered responses and independent Judge scores stored with each method-specific evaluation run, including the system-message settings described in Appendix D.3. The seven interior coordinates produce 22,148 new generations and the same number of Judge tasks per model–method condition, or 597,996 of each across all 27 conditions.

E.2 Aggregate control geometry

Let $s_1(\zeta)$ and $s_2(\zeta)$ denote the Judge support scores for values 1 and 2, and define

$$z(\zeta) = \frac{s_1(\zeta) - s_2(\zeta)}{4}.$$

Moving from ξ_1 to ξ_2 should make s_1 and z non-increasing and s_2 non-decreasing. We call a nine-point model–pair trajectory *zero-violation* when its source-averaged z curve has no adjacent increase $z(\zeta_{r+1}) > z(\zeta_r)$; equality is allowed. Aggregation otherwise follows Appendix D.5: original and reversed instances form one source cluster, clusters are averaged within each pair, and the 40 strict-Unseen pairs receive equal macro weight.

The endpoint span is directly related to the control metric in Appendix D.5:

$$z(0) - z(1) = 2\,\mathrm{BPC}.$$

It therefore restates the endpoints' control range. The seven interior samples add evidence about the intervening trajectory: how often it reverses direction, how large those reversals are, and how quality changes between the endpoints.

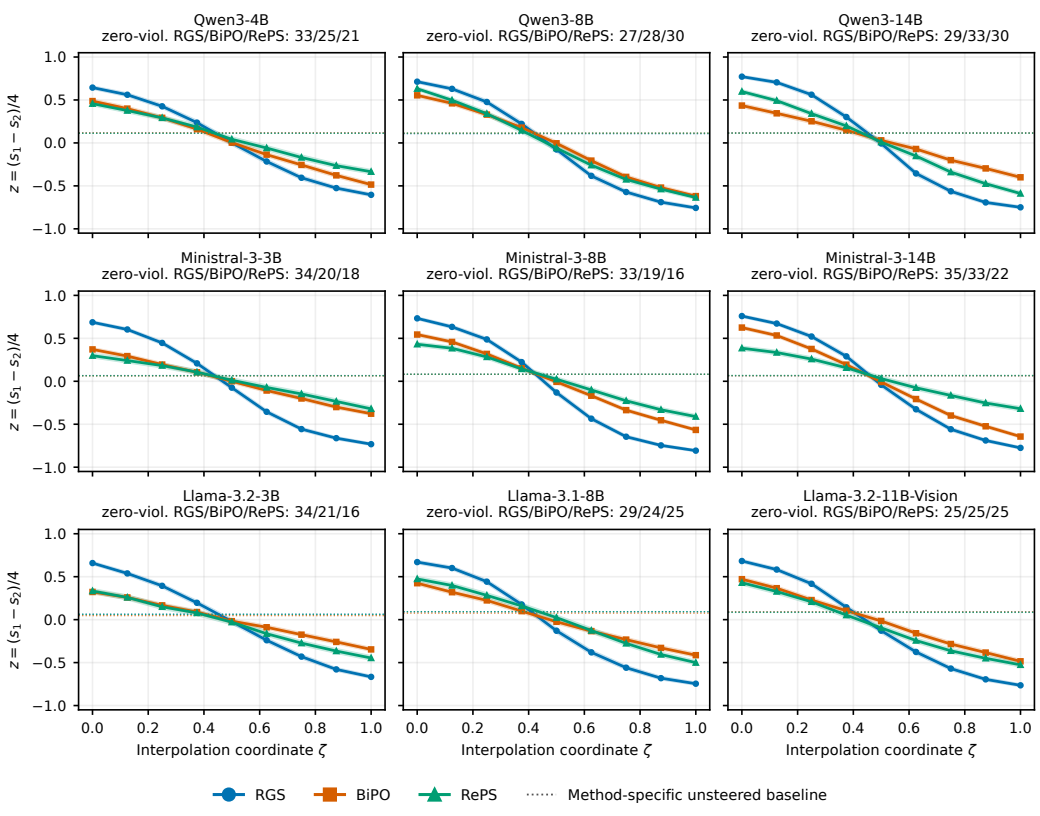


Figure 5: Sampled endpoint-interpolation trajectories on strict-Unseen value pairs. Each panel shows the equal-pair macro trajectory for one evaluated model at nine coordinates; shading gives 95% source-cluster-bootstrap confidence intervals, and the dotted lines mark method-specific unsteered baselines. All 27 model–method aggregate z trajectories move from value-1-favoring to value-2-favoring. RGS has the largest endpoint z-span in every model. Panel subtitles give the zero-violation counts out of 40 model–pair trajectories in RGS/BiPO/RePS order.

For the trajectory-level quality diagnostic, we first compute the quality difference between each interpolated response and the method-specific unsteered response to the same prompt. We average these differences within pairs and then equally across the 40 pairs to obtain each model's aggregate quality-change curve $\Delta Q(\zeta)$. Mean minimum ΔQ takes each model's minimum over the nine sampled coordinates and averages these minima across the nine models.

All 27 model–method aggregate trajectories have zero adjacent violations: z and s_1 are non-increasing at all eight sampled steps, while s_2 is non-decreasing. Here "zero adjacent violations" refers to the aggregate trajectories; the main text's "no reversals" statistic counts adjacent increases in pair-level z. The distinction between methods emerges at the model–pair level. Table 17 shows that RGS has the highest zero-violation rate, less than half as many reversal steps as either baseline, and smaller reversal magnitudes over its wider endpoint span.

Table 17: Sampled endpoint-interpolation geometry over the $9 \times 40 = 360$ strict-Unseen model–pair trajectories per method. Reversal magnitude gives the median, 90th percentile, and maximum positive adjacent increase in z. Mean minimum ΔQ first takes the lowest point of each model's nine-point aggregate quality curve relative to its unsteered baseline and then averages across the nine models.

Method	Zero-violation model–pairs	Reversal steps	Reversal magnitude med./p90/max	Mean z span	Mean min. ΔQ
RGS	**279/360 (77.5%)**	**95**	**.014/.042/.081**	**1.436**	-.158
BiPO	228/360 (63.3%)	217	.023/.080/.145	.952	-.116
RePS	203/360 (56.4%)	215	.024/.072/.157	.903	-.143

RGS attains the highest or tied-highest number of zero-violation pairs on seven of the nine models and the largest z-span on all nine. Its greater pair-level stability is therefore not obtained by restricting the observed control range. The two exceptions also show that span and local stability are distinct: RePS has the most zero-violation pairs on Qwen3-8B and BiPO has the most on Qwen3-14B, while both retain smaller spans than RGS on those models.

To measure how closely the sampled responses approach the unsteered levels of both support scores, we compute the minimum distance to that two-dimensional support state over the nine sampled coordinates. For this diagnostic, we first average support scores over source clusters within each pair and then equally over the 40 pairs, obtaining one aggregate trajectory for each model and method. Writing $s_{1,\mathrm{base}}, s_{2,\mathrm{base}}$ for the corresponding aggregate unsteered supports, this distance is

$$d_{2D} = \min_{\zeta \in \mathcal{Z}} \sqrt{\big(s_1(\zeta) - s_{1,\mathrm{base}}\big)^2 + \big(s_2(\zeta) - s_{2,\mathrm{base}}\big)^2}.$$

We estimate the crossing of $z_{\mathrm{base}} = (s_{1,\mathrm{base}} - s_{2,\mathrm{base}})/4$ by linear interpolation between adjacent sampled points of the aggregate z curve. We then average the crossing coordinates and d_{2D} values equally across the nine models. All three methods cross their own unsteered value preference at similar estimated mean coordinates (.433 for RGS, .418 for BiPO, and .420 for RePS). Across the nine sampled coordinates, their mean minimum d_{2D} values are .320, .217, and .166, respectively. Matching the value preference fixes only the support-score difference; it does not require both support scores to match their unsteered levels.

Across nine sampled coordinates on raw chords without norm matching, RGS combines a wider control range with fewer and smaller pair-level reversals. The cross-model mean of the minimum aggregate quality change is -0.158 for RGS, -0.116 for BiPO, and -0.143 for RePS (Table 17).

The complete pair-level trajectories are collected in Appendix G.1. They show the variation across value pairs behind the aggregate reversal statistics.

F Controlled mechanism studies and results

This appendix asks how the objective and endpoint parameterization affect behavioral movement. The primary component analysis is the complete zero-anchored (0/A/N) design on Qwen3-4B at seed 42, matching the nine configuration cells summarized in the main text. We then retain two earlier singleton-(A) studies as explicitly supplementary analyses, because their objective and (for the endpoint factorial) seed coverage differ from the primary design. Finally, we examine the evidence displacement geometry at selected steering configurations across all nine evaluated models.

The controlled component studies use Qwen3-4B at the layer and inference strength selected for RGS. Full-method comparisons use each method's own selected steering configuration and describe its resulting behavior. All comparisons use the common teacher-forced diagnostics in Appendix F.1. Appendix F.2 specifies the matched designs and Appendix F.3 describes the A- and N-critical challenge sets. The RGS component studies modify the fixed-weight smooth penalty surrogate in main-text Equation 9; this training objective does not guarantee per-record satisfaction of the hard constraints.

F.1 COMMON TEACHER-FORCED DIAGNOSTICS

All mechanism cells and the full RGS, RePS, and BiPO configurations are evaluated in the shared teacher-forced coordinate system of main-text Equations 1 and 3, regardless of their native training objectives. Within each diagnostic cell, η denotes its fixed method-specific inference strength. For RGS, we evaluate main-text Equation 3 at $\alpha = \eta$; for RePS, the independently trained value-specific parameters (w_d, b_d) give the positive intervention $h' = h + (\eta + b_d)w_d$; for BiPO, we use $h' = h + \iota_d \eta u_p$, where u_p is the local vector for the current value pair and $\iota_d \in \{-1, +1\}$ selects the sign associated with direction d. For RePS and BiPO, $\Delta_X^{(d)}(\eta)$ compares the resulting score with the unsteered score. These coordinates measure intended, alternative, and natural evidence displacements in the standard cells. A source–direction combination constitutes one directional record. Under direction d, I is the intended response and A the alternative response. In the standard cells, N is the unconditioned model-native natural response; the challenge cells instead use the audited substantive background-hybrid instantiation defined in Appendix F.3. For each record, we define

$$\begin{aligned} m_0^{(d)}(\eta) &= \Delta_I^{(d)}(\eta), & m_A^{(d)}(\eta) &= \Delta_I^{(d)}(\eta) - \Delta_A^{(d)}(\eta), \\ m_N^{(d)}(\eta) &= \Delta_I^{(d)}(\eta) - \Delta_N^{(d)}(\eta), & m_{\text{worst}}^{(d)}(\eta) &= \min\Big(m_0^{(d)}(\eta), m_A^{(d)}(\eta), m_N^{(d)}(\eta)\Big). \end{aligned} \tag{F.1}$$

The margin subscript identifies the competitor: zero uplift, the alternative response, or the natural response. Thus m_0, m_A, and m_N are the raw diagnostic margins associated with (G), (S_A), and (S_N) in the main text. Their corresponding constraint slacks are $m_0 - \rho$, $m_A - \rho$, and $m_N - \rho$. For (L), $\Delta_A - \epsilon$ is the violation and $\epsilon - \Delta_A$ is the slack. The worst-case target margin m_{worst} is a hard minimum and differs from the smooth training margin m_τ defined in Section 3.3. Because the minimum is taken per record, the reported mean of m_{worst} need not equal the minimum of the three separately averaged margins. Negative-I is the fraction of directional records with $\Delta_I^{(d)}(\eta) < 0$. To identify which competitor exerts the largest pressure in the same coordinate system, we also define the diagnostic responsibilities

$$\Big(\chi_0^{(d)}(\eta), \chi_A^{(d)}(\eta), \chi_N^{(d)}(\eta)\Big) = \text{softmax}\left(\frac{\Big(0, \Delta_A^{(d)}(\eta), \Delta_N^{(d)}(\eta)\Big)}{\tau}\right), \qquad \tau = 0.1. \tag{F.2}$$

The largest component identifies the hardest current competitor. These evaluation-time weights describe the relative competitor pressure implied by the shared LSE diagnostic; component effects are estimated separately through the matched contrasts below. Reported tables suppress the fixed (η) argument.

For the standard five-pair Validation diagnostics, each training-seed cell contains 50 semantic sources per value pair, each scored under both steering directions, for 100 directional records per pair. For each source and direction, each family score is first averaged over its three fixed responses using the normalized family weights; the three displacements and record-level margins are then formed from those family means. This differs from the RGS validation-monitoring loss in Table 5, which is excluded from gradient optimization and enumerates all $3^3 = 27$ response triplets. Record-level diagnostics are averaged with equal weight over directions and sources within each pair; the five pair estimates are then macro-averaged equally. This aggregation applies to the non-challenge RGS cells, the full-method configurations, and the RePS bridge. The A- and N-critical cells use the same record-level definitions but the paired-seed aggregation specified in Appendix F.3.

F.2 CONTROLLED DESIGNS AND ESTIMANDS

Within-RGS mechanism cells use Qwen3-4B, zero-based layer $\ell = 20$ at the post-block residual stream, RGS training intervention coefficient $\alpha_{\text{train}} = 1$, inference strength $\eta = 1.5$, and the joint epoch-40 checkpoint. They share the five training value pairs and 400/50 source splits. All primary cells use seed 42; the earlier singleton-A endpoint factorial is additionally repeated with seeds 43 and 44. Table 18 maps the short names used in the analysis to controlled objective variants. Across matched RGS cells, all unlisted settings remain fixed, including $\rho, \beta, \tau, \epsilon$, the separation and energy weights, the hard budget $B = 30$, the optimizer, and the training data. Energy is computed over both endpoints: for a signed axis $\boldsymbol{u}_1 = \boldsymbol{u}, \boldsymbol{u}_2 = -\boldsymbol{u}$, it is $2\|\boldsymbol{u}\|_2^2$, preserving the endpoint-energy convention across parameterizations.

Table 18: Mechanism cell dictionary. P uses the singleton competitor term Δ_A, whereas $0AN = \mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$ is the full smoothed competitor maximum. The separation and energy coefficients remain fixed across cells.

Cell	Endpoint parameterization	Competitor construction	Enhance weight	Leakage weight
P-S-F0	Signed axis	Δ_A	0	0
P-I-F0	Separate vectors	Δ_A	0	0
P-S-F12	Signed axis	Δ_A	0.5	0.2
P-I-F12	Separate vectors	Δ_A	0.5	0.2
P-I-F1	Separate vectors	Δ_A	0.5	0
P-I-F2	Separate vectors	Δ_A	0	0.2
0A-I-F0	Separate vectors	$\mathrm{LSE}_\tau(0, \Delta_A)$	0	0
0AN-I-F0	Separate vectors	$\mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$	0	0
0AN-S-F0	Signed axis	$\mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$	0	0
0AN-I-F1	Separate vectors	$\mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$	0.5	0
0AN-I-F2	Separate vectors	$\mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$	0	0.2
0AN-S-F12	Signed axis	$\mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$	0.5	0.2
0AN-I-F12	Separate vectors	$\mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$	0.5	0.2
0N-I-F1	Separate vectors	$\mathrm{LSE}_\tau(0, \Delta_N)$	0.5	0
0A-I-F12	Separate vectors	$\mathrm{LSE}_\tau(0, \Delta_A)$	0.5	0.2

Each cell label encodes the competitor members, endpoint parameterization, and enabled terms. The prefix `P` denotes the singleton pairwise competitor term Δ_A; `0A`, `0N`, and `0AN` denote membership in the zero-anchored LogSumExp competitor maximum. The components `S` and `I` distinguish endpoint parameterizations: `S` imposes the signed-axis constraint $\boldsymbol{u}_C = -\boldsymbol{u}_T$, whereas `I` removes that constraint and learns two endpoints. Within the suffixes, `F0` means that both toggled components are off; among these two components, `F1` enables only enhance, `F2` enables only leakage, and `F12` marks both terms. The enhance and leakage weights of 0.5 and 0.2 are the coefficients used in training. For full-objective cells, the unit-weight enhance convention in main-text Equation 9 is obtained by a common positive rescaling of the loss and its remaining coefficients, as detailed in Appendix A. P-I-F0 and P-I-F12 are shared by the enhance-by-leakage and endpoint-by-bundle designs at seed 42; 0AN-I-F12 is the full RGS objective.

Primary complete-$0/A/N$ design. The primary appendix table reports the nine seed-42 cells used by the revised main-text ablation. The generation and Judge metrics for all nine rows use seed 42. The signed no-term corner $\mathrm{0AN-S-F0}$ was completed post hoc; seeds 43 and 44 provide common-score replications only and are not additional generation/Judge replications. For strict-Unseen generation/Judge, point estimates macro-average the 40 value pairs. Within each pair, original and reversed prompt instances sharing a source cluster are resampled jointly; 10,000 multinomial source-cluster bootstrap replicates use shared weights across all cells, and pair means are then averaged equally. The 2.5th and 97.5th percentiles form the intervals. These intervals condition on seed 42 and quantify source-cluster sampling only; they do not cover training-seed uncertainty.

Table 19: Primary nine-cell ablation on Qwen3-4B (seed 42). The first four columns are common Validation teacher-forced means; BPC and Q_s are strict-Unseen generation/Judge means. m_{worst} is the mean of record-level minima, not the minimum of the displayed cell means.

Cell	Δ_I	Δ_A	Δ_N	m_{worst}	BPC	Q_s
P-I-F0	-12.651	-22.070	-18.467	-12.651	.354	2.367
0A-I-F0	+1.307	-1.935	-.938	+1.074	.587	4.351
0AN-I-F0	+.880	-2.044	-1.713	+.792	.623	4.043
0AN-I-F1	+2.279	+1.706	+1.280	+.518	.454	4.447
0AN-I-F2	+.760	-1.391	-.963	+.713	.596	4.306
0AN-I-F12	+1.771	-.809	-.242	+1.312	**.624**	4.333
P-S-F0	-3.208	-7.870	-5.952	-3.210	.489	3.424
0AN-S-F0	+1.011	-.576	-.286	+.743	.456	4.247
0AN-S-F12	+1.692	-.193	+.393	+.995	.445	3.889

The complete $0/A/N$ endpoint-by-bundle design uses $0\mathrm{AN}{-}\mathrm{S}{-}\mathrm{F0}$, $0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F0}$, $0\mathrm{AN}{-}\mathrm{S}{-}\mathrm{F12}$, and $0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F12}$. For any outcome Y, its interaction is

$$\begin{aligned}\mathcal{I}^{0AN}_{\mathrm{endpoint}\times\mathrm{bundle}} = &\big[Y(0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F12}) - Y(0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F0})\big] \\ &- \big[Y(0\mathrm{AN}{-}\mathrm{S}{-}\mathrm{F12}) - Y(0\mathrm{AN}{-}\mathrm{S}{-}\mathrm{F0})\big].\end{aligned} \tag{F.3}$$

The strict-Unseen BPC endpoint gains are $+0.167$ without the bundle and $+0.179$ with it; their interaction is $+0.012\,[-0.007, +0.030]$. These intervals use fixed-seed source-cluster bootstrap and do not quantify training-seed uncertainty. Without the bundle, separate endpoints lower Δ_I by 0.131 and Q_s by 0.204; with the bundle they increase Δ_I by 0.080, m_{worst} by 0.317, and Q_s by 0.444.

The complete-$0/A/N$ enhancement/leakage comparison fixes separate endpoints and uses $0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F0}$, $0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F1}$, $0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F2}$, and $0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F12}$. Relative to $0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F0}$, enhancement alone raises Δ_I from 0.880 to 2.279 but also raises alternative and natural uplift and lowers BPC from 0.623 to 0.454. Leakage alone does not improve intended uplift, worst-case target margin, or BPC. Adding leakage to enhancement lowers Δ_A from $+1.706$ to -0.809 and restores BPC to 0.624. These are cell-level point estimates in the primary seed-42 design; the registered source-bootstrap intervals below belong to the explicitly marked supplementary singleton-A contrast.

Supplementary singleton-A designs. The following two studies predate the complete-(0/A/N) analysis and are retained as supplementary matched analyses. They use the singleton competitor term Δ_A, so their effects must not be read as estimates of the complete-(0/A/N) design. The endpoint factorial additionally uses training seeds 42–44.

The supplementary endpoint study is a parameterization-by-term-bundle 2×2: P-S-F0, P-I-F0, P-S-F12, and P-I-F12. It is repeated with training seeds 42, 43, and 44. It estimates, within this matched design, the interaction between endpoint freedom and the jointly enabled enhance/leakage bundle; it does not separately identify the enhance term, the leakage term, zero, or natural-response membership. The supplementary enhance and leakage terms are separated in a second four-cell design that fixes separately parameterized endpoints and Δ_A: P-I-F0, P-I-F1, P-I-F2, and P-I-F12.

The competitor ladder fixes separately parameterized endpoints and disables the enhance and leakage terms while moving from Δ_A to $\mathrm{LSE}_\tau(0, \Delta_A)$ and then to $\mathrm{LSE}_\tau(0, \Delta_A, \Delta_N)$. Since $\mathrm{LSE}_\tau(\Delta_A) = \Delta_A$, these transitions add zero and then the natural-response displacement to the same LogSumExp competitor construction. A separate full-objective comparison, 0AN-I-F12 versus 0AN-S-F12, tests endpoint parameterization with the full RGS smooth penalty surrogate.

For any outcome Y, the two supplementary singleton-A interaction estimands are

$$\begin{aligned}\mathcal{I}_{\mathrm{enh}\times\mathrm{leak}} = &\big[Y(\mathrm{P}{-}\mathrm{I}{-}\mathrm{F12}) - Y(\mathrm{P}{-}\mathrm{I}{-}\mathrm{F1})\big] \\ &- \big[Y(\mathrm{P}{-}\mathrm{I}{-}\mathrm{F2}) - Y(\mathrm{P}{-}\mathrm{I}{-}\mathrm{F0})\big],\end{aligned} \tag{F.4}$$

$$\begin{aligned}\mathcal{I}_{\mathrm{endpoint}\times\mathrm{bundle}} = &\big[Y(\mathrm{P}{-}\mathrm{I}{-}\mathrm{F12}) - Y(\mathrm{P}{-}\mathrm{I}{-}\mathrm{F0})\big] \\ &- \big[Y(\mathrm{P}{-}\mathrm{S}{-}\mathrm{F12}) - Y(\mathrm{P}{-}\mathrm{S}{-}\mathrm{F0})\big].\end{aligned} \tag{F.5}$$

The first uses one fixed training seed and paired source-cluster bootstrap uncertainty. The second uses the source-cluster bootstrap for this earlier singleton-A design; its seed-specific point estimates are reported below, while the three-seed coverage is descriptive rather than a new uncertainty estimate. The full-objective endpoint contrast is $Y(0\mathrm{AN}{-}\mathrm{I}{-}\mathrm{F12}) - Y(0\mathrm{AN}{-}\mathrm{S}{-}\mathrm{F12})$.

Full RGS, native BiPO, and native RePS are additionally evaluated with the common diagnostics in Appendix F.1 at their own selected steering configurations. These comparisons describe the behavioral regime reached by each full method. The RePS-A bridge is a separate matched contrast at the steering configuration selected for native RePS. We refer to the native natural-response training configuration as RePS-N. The bridge retains the rank-one parameterization, layer, inference strength, positive and nulling interventions, and optimizer, while replacing only the unconditioned training response family with an explicit alternative-behavior response.

F.3 A- AND N-CRITICAL CHALLENGE SETS

Member-specific effects use independently constructed, disjoint A-critical (alternative-critical) and N-critical (background-critical) challenge sets over the 40 strict-Unseen value combinations. In

this challenge-set subsection, N is an audited substantive background-hybrid response family constructed to stress constraint (S_N); it is not the unconditioned model-native natural-response family used in RGS training. Candidate source identifiers come from the open-source CoopValue records in these 40 pairs; exclusions and the resulting source order are fixed by the construction code and frozen manifests. Eligible sources follow a fixed deterministic order within each value pair and are then interleaved round-robin across pairs. The first 500 sources form the N-critical construction tranche and the next 500 form the disjoint A-critical construction tranche. This ordering is fixed before scenario generation, and no RGS output enters candidate ordering or membership.

GPT-5.4 rewrites each candidate into a scenario seed that fixes three executable policies—one value-aligned policy for each of the two values and one substantive background hybrid—and three response paraphrases per policy. Original and reversed prompt instances reuse the same policies and responses, changing only the actor preference assignment. A separate GPT-5.4 audit and a DeepSeek-V4-Flash audit score feasibility and semantic competitiveness for both orientations on 1–5 scales. Every policy must receive a feasibility score of at least 4. For A-critical candidates, the two value-aligned policies differ in competitiveness by at most one point and background competitiveness is at least two points below both value-aligned policies; for N-critical candidates, background competitiveness is at least two points above both value-aligned policies. Deterministic checks and both audits must confirm the required ordering and comparable response quality. The first 200 consensus seeds in frozen order are retained for each set before any RGS scoring. Complete construction and audit messages are defined in the construction code; selected source identifiers and construction provenance are retained in the frozen manifests.

The matched comparisons reported here reuse these benchmark files unchanged with the Qwen3-4B $\ell = 20$, $\eta = 1.5$, epoch-40 cells specified in Appendix F.2. On the A-critical set, 0AN-I-F1 is compared with 0N-I-F1, deleting only A; both cells retain the enhance term and disable the leakage term. Leakage is disabled because its penalty depends on Δ_A, so retaining it would preserve A-dependent loss pressure after removing A from the LogSumExp competitor maximum. On the N-critical set, 0AN-I-F12 is compared with 0A-I-F12, deleting only N while retaining both terms because leakage does not depend on N. Each challenge set contains 200 paired scenario seeds, 400 original/reversed prompts, and 800 directional records per cell. The three paraphrases in each fixed answer family are averaged before the record-level diagnostics in Appendix F.1 are computed; these answers are teacher-forced targets, not open-ended evaluation generations. Each paired seed groups its two orientations and two steering directions as one resampling unit. Point estimates first average the four directional records within a seed and then weight the 200 seeds equally. The 95% confidence intervals use 10,000 paired-seed bootstrap replicates. Each replicate samples 200 seed identifiers with replacement, jointly carries their matched cells and all four directional records, and recomputes the equal-seed mean; the 2.5th and 97.5th percentiles form the interval endpoints.

These teacher-forced contrasts estimate Qwen3-4B member effects in two targeted challenge distributions: A-critical enables enhance and disables leakage, while N-critical retains both terms, with full RGS on its 0AN side. Each effect is interpreted within its corresponding objective and challenge distribution. The targeted sampling does not estimate the prevalence of these failure modes in natural CoopValue data.

F.4 Mechanism results

F.4.1 Pairwise success does not identify behavioral movement

We first test whether a large intended–alternative margin can coexist with a decrease in the intended response score. To examine the underidentification discussed in Section 3.1, we use matched pairwise probes implemented with the RGS parameterization and training pipeline. Their behavioral data-fit depends only on the intended–alternative margin, with no enhance or leakage term; all other fixed regularization remains unchanged. All cells are evaluated with the common teacher-forced diagnostics defined in Appendix F.1. As Table 20 shows, both signed-axis and separately parameterized probes attain large positive $m_A = \Delta_I - \Delta_A$ margins but have negative mean Δ_I. The separately parameterized probe reaches a stronger common-suppression regime. Hence, relative success is compatible with failure relative to the unsteered model.

At their selected steering configurations, the same diagnostic distinguishes the full methods descriptively. BiPO has the pairwise underidentification established by the common-shift argument in

Table 20: Common-score diagnostics on Qwen3-4B. In the first two matched pairwise probes, implemented with the RGS parameterization and training pipeline, the behavioral data-fit depends only on $\Delta_I - \Delta_A$; the final three rows are full methods at their own selected steering configurations and are descriptive, not component contrasts. All rows use seed 42. Negative-I is the percentage of directional records with $\Delta_I < 0$.

Cell	Δ_I	Δ_A	m_A	Δ_N	m_{worst}	Neg.-I
Pairwise signed	-3.208	-7.870	4.663	-5.952	-3.210	83.0%
Pairwise separate vectors	-12.651	-22.070	9.418	-18.467	-12.651	100.0%
BiPO	0.372	-1.283	1.656	-0.803	0.328	25.8%
RePS	2.632	2.008	0.625	1.583	0.578	0.0%
RGS	1.771	-0.809	2.580	-0.242	1.312	3.4%

Section 3.1 in its native summed-score coordinates. Table 20 instead measures its observed behavior with the shared normalized, discounted scores: BiPO primarily separates intended and alternative responses by suppressing the alternative response, and its intended evidence displacement is negative on 25.8% of records. RePS increases the intended score but also increases alternative and natural scores on $100\%/100\%$ of records. Because native RePS uses an unconditioned response in an asymmetric scaled-SimPO objective, it is not a pure $\Delta_I - \Delta_A$ objective and is not covered by that argument. Its row reveals a distinct non-selective-uplift regime in the common coordinates. Among these Qwen3-4B full-method diagnostics, RGS combines positive intended movement with the largest worst-case target margin.

Explicit alternative behavior in RePS. This contrast tests whether training against explicit alternative responses makes RePS more selective. At the steering configuration selected for native RePS, we replace the unconditioned training response family with explicit alternative-behavior responses while holding the rank-one parameterization, positive and nulling interventions, optimizer, layer, and inference strength fixed. Reporting RePS-A minus native RePS-N, this matched bridge improves the teacher-forced worst-case target margin by $0.093\,[0.073, 0.114]$, Seen BPC by $0.080\,[0.029, 0.133]$, and strict-Unseen BPC by $0.085\,[0.068, 0.103]$; Order-preserving rate and Target-value Induction Rate also improve. Intended uplift nevertheless decreases by 0.222, strict-Unseen mean steered quality decreases by 0.432, and unclear/refusal increases by 0.048. Alternative and natural uplift remain positive on 98.8% and 98.4% of Validation records. Explicit alternative behavior therefore shifts the RePS regime but only partially improves the common behavioral-control diagnostics.

F.4.2 RGS COMPONENT MECHANISMS

The enhance/leakage, endpoint, and competitor-ladder cells pair Validation teacher-forced diagnostics with generation on the 40 strict-Unseen value combinations; all BPC effects below are strict-Unseen. The A/N member tests use the separate fixed teacher-forced challenge sets in Appendix F.3.

Supplementary singleton-A enhance and leakage control. This supplementary singleton-A study tests how encouraging intended uplift and penalizing alternative uplift affect control. We hold separately parameterized endpoints and the singleton competitor term Δ_A fixed and cross the enhance term with the alternative leakage term. Table 21 reports the teacher-forced coordinates and strict-Unseen generation behavior.

Table 21: Qwen3-4B seed-42 enhance-by-leakage study with separately parameterized endpoints and the singleton competitor term Δ_A. BPC and U/R are strict-Unseen; the remaining columns use the common Validation diagnostics.

Enabled term	Δ_I	Δ_A	Δ_N	m_{worst}	BPC	U/R
None	-12.651	-22.070	-18.467	-12.651	0.354	0.387
Enhance	2.270	1.630	1.248	0.570	0.463	0.013
Leakage	-12.190	-21.413	-17.923	-12.190	0.370	0.359
Both	1.716	-1.166	-0.402	1.296	0.616	0.009

In this seed-42, alternative-only Δ_A design, the enhance term alone reverses the sign of Δ_I and reduces unclear/refusal from 0.387 to 0.013, but it also uplifts A/N on $99.4\%/99.0\%$ of Validation records. The leakage term alone leaves the joint-suppression solution largely unchanged. Conditional on the enhance term, however, leakage reduces Δ_A by 2.796, increases m_{worst} by 0.726, and improves strict-Unseen BPC by $0.153\,[0.138, 0.168]$. The enhance-by-leakage BPC interaction is $0.136\,[0.119, 0.154]$. The effect is therefore conditional rather than merely additive: leakage improves alternative-behavior control when the enhance term is present but leaves the joint-suppression solution largely unchanged without it. The intervals describe source-cluster resampling uncertainty conditional on this fixed training run.

Supplementary singleton-A constrained endpoint flexibility. This supplementary experiment asks when giving the two endpoints separate parameters improves control over using opposite signs of one vector. The supplementary factorial experiment crosses endpoint parameterization with the joint enhance/leakage bundle while keeping the singleton competitor term Δ_A fixed. The separately-parameterized-minus-signed-axis BPC effect changes from $-0.138\,[-0.152, -0.123]$ without the bundle to $+0.182\,[0.169, 0.195]$ with it; the interaction is $+0.320\,[0.301, 0.338]$. The same crossover appears in Order-preserving rate, Target-value Induction Rate, mean steered quality, and unclear/refusal. Under the full RGS smooth penalty surrogate, separately parameterized endpoints still improve BPC by $0.179\,[0.165, 0.194]$ and the teacher-forced worst-case target margin by 0.317. The benefit of endpoint freedom therefore depends on the objective terms that guide it. Complete cell means and seed-specific estimates follow below.

Table 22 gives the complete cell means for the supplementary endpoint-by-term-bundle design.

Table 22: Strict-Unseen means over Qwen3-4B seeds 42–44. The design crosses endpoint parameterization with the joint enhance/leakage bundle while fixing the singleton competitor term Δ_A. Steered Q is $(Q^{(1)} + Q^{(2)})/2$; q in Equation D.3 is the corresponding baseline-relative change.

Endpoint	Enh./leak.	BPC	OrderPres	Induction	Steered Q	U/R
Signed	Off	0.491	0.599	0.739	3.431	0.082
Separate vectors	Off	0.354	0.405	0.504	2.362	0.385
Signed	On	0.433	0.616	0.642	3.853	0.050
Separate vectors	On	**0.615**	**0.738**	**0.767**	**4.329**	**0.009**

The supplementary seed-specific BPC interaction estimates are $0.319/0.316/0.324$. In the separate full-objective comparison, separately parameterized endpoints improve mean steered quality by $0.444\,[0.416, 0.472]$, in addition to the BPC and effects on the worst-case target margin reported above.

Competitor structure (primary ladder and supplementary bridge). We test how grounding the objective at zero uplift, and then adding natural-response competition, changes the learned behavior. The seed-42 competitor ladder makes the matched transitions explicit: $\mathrm{P{-}I{-}F0} \rightarrow \mathrm{0A{-}I{-}F0} \rightarrow \mathrm{0AN{-}I{-}F0}$. The first transition replaces Δ_A with $\mathrm{LSE}_\tau(0, \Delta_A)$, reverses mean intended movement from -12.651 to $+1.307$, improves strict-Unseen BPC by $0.233\,[0.218, 0.248]$, and reduces unclear/refusal by 0.377. Because $\Delta_A = \mathrm{LSE}_\tau(\Delta_A)$, this is the effect of adding the zero competitor under the fixed no-enhance/no-leakage objective. The second transition adds only N to the same LogSumExp competitor maximum under the no-enhance/no-leakage context; it increases mean BPC by $0.036\,[0.024, 0.048]$ while reducing mean steered quality by 0.308.

As a supplementary singleton-A contrast, with enhance and leakage enabled, the separate seed-42 contrast $\mathrm{0AN{-}I{-}F12} - \mathrm{P{-}I{-}F12}$ changes the singleton A competitor into the full $0/A/N$ smoothed competitor maximum. Its BPC estimate is small and positive, $0.008\,[0.000, 0.016]$. This contrast estimates the joint contribution of zero and N with enhance and leakage enabled; the average BPC gain is much smaller than in the no-term ladder.

The challenge sets ask whether including a competitor protects separation from it on scenarios designed to make it competitive. We compare matched objectives with and without the targeted member. In the A-critical set, the 0AN-I-F1 comparator assigns evaluation-time mean responsibility $\chi_A = 0.803$ to A; adding A improves m_A by $0.0214\,[0.0178, 0.0250]$ and the worst-case target margin by $0.0196\,[0.0162, 0.0230]$. In the N-critical set, the 0AN-I-F12 comparator assigns

$\chi_N = 0.626$ to N; adding N improves m_N by $0.0134\,[0.0096, 0.0172]$ and the worst-case target margin by $0.0214\,[0.0166, 0.0264]$. These are boundary reallocations rather than simultaneous improvements of every margin: the A-member contrast changes m_0 by $-0.1412\,[-0.1515, -0.1311]$, and the N-member contrast changes m_A by $-0.0044\,[-0.0077, -0.0010]$.

F.4.3 Cross-model behavioral geometry at selected steering configurations

We next ask how intended uplift and competitor suppression vary across models at their selected steering configurations. We visualize the directional-record distributions in the same common teacher-forced coordinates for all nine evaluated models. For each model and method, the diagnostic contains the five fixed training-pair identities, 50 Validation sources per pair, and both steering directions, for 500 directional records after family-first aggregation. The three formal methods use their own selected steering configurations and native interventions. The pairwise signed and separate-vector probes instead use the corresponding RGS layer and strength and are geometry probes rather than independently selected formal baselines.

Qwen3-4B uses the same system-message scoring records as Table 20. The nine models use the canonical decision-scenario system message quoted in Appendix D.3. Within each model, all five configurations share the scoring prompt, and the reference scale below uses the matching unsteered scores. These are within-model comparisons; the scoring prompt is identical across all nine models.

Raw displacement scales differ substantially across models. We first retain the common absolute threshold $\rho = 0.8$, because it is the margin target specified by RGS and therefore provides the primary fixed-threshold audit. For any diagnostic threshold r, define

$$J^{(d)}(r) = \mathbf{1}\Big\{m_\tau^{(d)} \geq r, \quad \Delta_A^{(d)} \leq \epsilon\Big\}, \tag{F.6}$$

and report its mean over directional records. Because LogSumExp upper-bounds the hard competitor maximum, $J^{(d)}(r) = 1$ is a sufficient but not necessary certificate that the family-first aggregated diagnostic record satisfies the four behavioral inequalities with ρ replaced by r. This certificate applies to the post-training family means; individual response triplets and the intervention-energy budget are outside this diagnostic.

At the original $\rho = 0.8$, RGS passes this smoothed diagnostic on 51.8–57.8% of Qwen records and 0–8.0% of Ministral/Llama records (Table 23). Thus RGS has low absolute diagnostic rates on six models, despite ranking highest or tied-highest in each model. Since J is a sufficient criterion, records outside it may still satisfy the hard behavioral inequalities.

Figure 6 shows the complete Qwen3-4B distribution corresponding to the regimes in Table 20. RGS combines positive mean intended movement with negative mean alternative movement and has the largest smoothed-control diagnostic rate. BiPO has weaker intended uplift, the pairwise probes suppress both scores, and RePS raises both. The pale-green region shows the two-dimensional conditions for the grounding, competition, and leakage inequalities at the fixed threshold: $\Delta_I \geq 0.8$, $\Delta_A \leq \Delta_I - 0.8$, and $\Delta_A \leq \epsilon$, with the observed Δ_N omitted. The panel percentage reports the mean of $J^{(d)}(0.8)$ in Equation F.6, which also uses the observed Δ_N. KDE contours show within-panel density shape, not confidence regions.

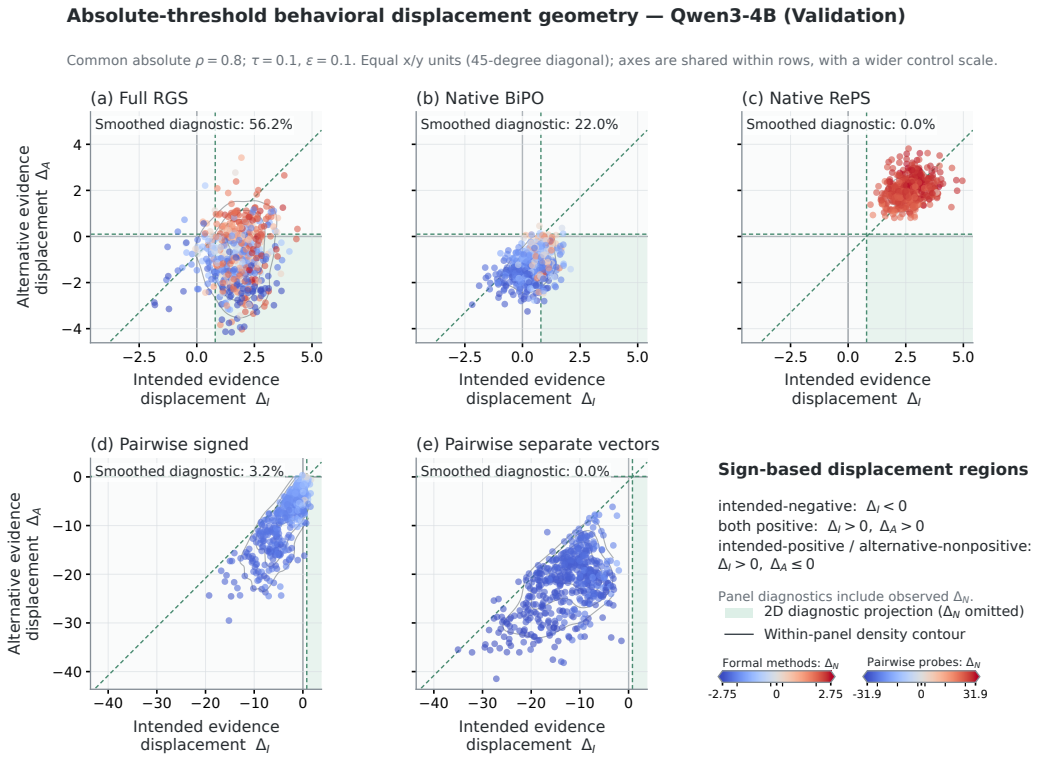


Figure 6: Fixed-threshold teacher-forced behavioral displacement geometry for Qwen3-4B. Formal methods use their own selected steering configurations; both pairwise probes use the RGS layer and strength. Each point represents one Validation source under one steering direction after family-first aggregation; the axes show raw Δ_I and Δ_A, and color encodes raw Δ_N (500 records per method). The pale-green region projects $\Delta_I \geq 0.8$, $\Delta_A \leq \Delta_I - 0.8$, and $\Delta_A \leq \epsilon$; the natural-response separation condition and intervention-energy budget are omitted. Panel percentages report the mean of $J^{(d)}(0.8)$, which uses the observed Δ_N. Formal methods and pairwise probes use separate row-wise axis ranges, so pixel spans are not comparable across rows. Point colors likewise use separate model-specific pooled symmetric-log scales; triangular colorbar tips mark clipping at the row-specific 98th-percentile absolute Δ_N, so hues are not comparable across rows or models.

The corresponding distributions for the remaining eight models appear in Appendix G.2, using the same absolute threshold and plotting conventions.

Sensitivity to the reference-score scale. We check whether method rankings change when the diagnostic uses a fixed absolute margin threshold or a reference-scaled threshold. The common $\rho = 0.8$ criterion asks whether every model reaches the same absolute margin, but the unsteered family-score scale differs substantially across models. Let $\bar{S}_{mpiF}(0)$ denote the unsteered family-first score for model m, pair p, source i, and family $F \in \{T, C, N\}$. We define a method-blind reference scale and a model-relative diagnostic threshold by

$$s_m = \operatorname*{median}_{p} \operatorname*{median}_{i,\,(F,F')\in\{(T,C),(T,N),(C,N)\}} \left|\bar{S}_{mpiF}(0) - \bar{S}_{mpiF'}(0)\right|, \qquad r_m(\kappa) = \kappa s_m. \quad \text{(F.7)}$$

The scale uses only unsteered reference scores and is shared by all methods for model m; it is not chosen from method outcomes. The fixed threshold corresponds to $\kappa_{0.8,m} = 0.8/s_m$: it is approximately 0.76–1.23 reference units for the three Qwen models but 2.54–4.38 units for the Ministral and Llama models. A common absolute threshold can therefore induce cross-model floor effects even though it remains the criterion directly tied to the RGS specification.

We consequently retain $r = \rho = 0.8$ as the primary diagnostic and additionally report the post-hoc model-relative view $r = r_m(\kappa)$. This substitution changes only the diagnostic threshold: the plotted

$\Delta_I, \Delta_A, \Delta_N$, $\tau = 0.1$, $\epsilon = 0.1$, and all selected steering configurations remain unchanged. Because τ and ϵ are not rescaled, this is a reference-scaled threshold sensitivity rather than a complete dimensionless normalization. A threshold below $\rho = 0.8$ gives a weaker diagnostic and cannot certify the original margin; a threshold at or above ρ is at least as strict, so $J^{(d)}(r_m(\kappa)) = 1$ implies $J^{(d)}(\rho) = 1$.

Figure 7 gives one representative model-relative scatter view. The remaining cross-model comparison is reported compactly in Table 23 and the threshold-sensitivity curves below.

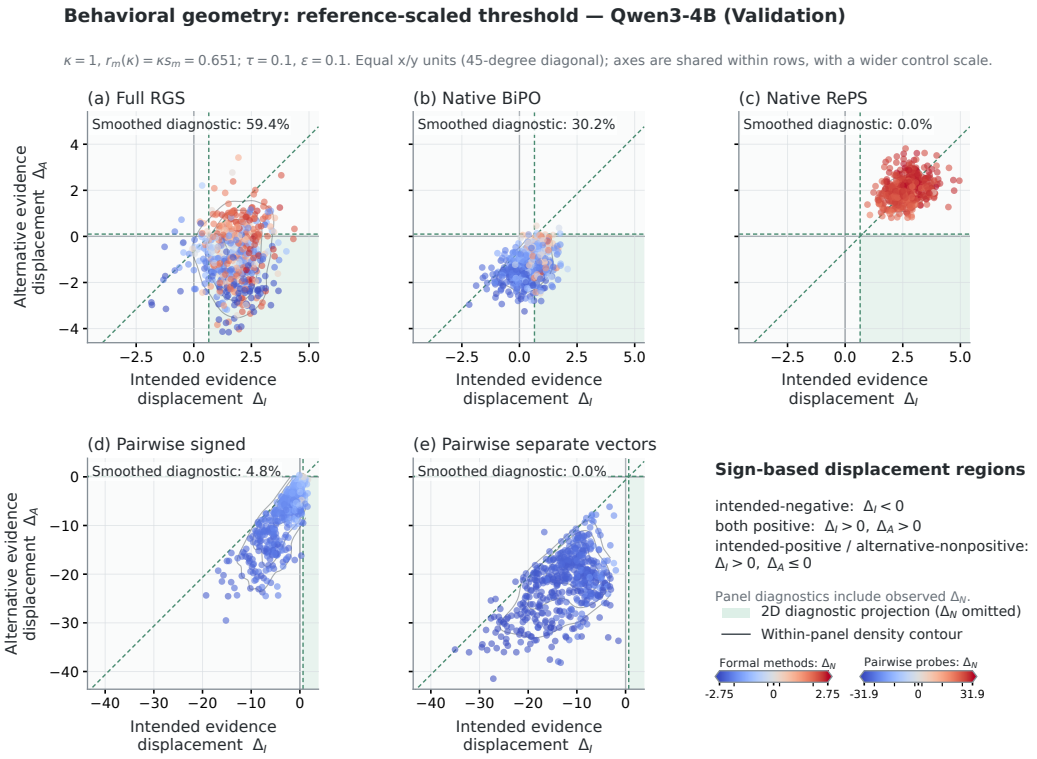


Figure 7: Model-relative (reference-scaled-threshold) view of Qwen3-4B geometry at the selected steering configurations for $\kappa = 1$, where $r_m(1) = s_m = 0.651$. The underlying records, raw coordinate axes, KDE contours, and Δ_N color scale are identical to Figure 6; only the diagnostic threshold, pale-green projection, and panel percentages change. Here the reference-scaled threshold is less strict than $\rho = 0.8$.

Table 23 reports all nine fixed-model point estimates under both threshold views. The reference-scaled view tests how the comparison changes when the margin is expressed relative to each model's unsteered score scale. At $\kappa = 1$, RGS has the highest smoothed-control diagnostic point estimate for every fixed model. All 18 within-model RGS-minus-BiPO and RGS-minus-RePS differences are positive, and their pointwise paired percentile 95% intervals at $\kappa = 1$ exclude zero. Under the common absolute $\rho = 0.8$, RGS is strictly highest for eight models and tied at zero with both formal baselines for Ministral-3-3B.

Table 23: Post-hoc smoothed-control diagnostic point estimates for the three formal methods. "Fixed" uses the common threshold $\rho = 0.8$; "Ref.-scaled" uses $r_m(1) = s_m$. Each cell averages 500 directional records over five fixed pairs. Bold marks the highest value within a model and threshold view; all three fixed-threshold entries for Ministral-3-3B are tied.

	Fixed $\rho = 0.8$			Ref.-scaled $\kappa = 1$		
Model	RGS	BiPO	RePS	RGS	BiPO	RePS
Qwen3-4B	**.562**	.220	.000	**.594**	.302	.000
Qwen3-8B	**.518**	.170	.272	**.370**	.088	.250
Qwen3-14B	**.578**	.246	.128	**.624**	.274	.132
Ministral-3-3B	**.000**	**.000**	**.000**	**.278**	.040	.032
Ministral-3-8B	**.018**	.000	.006	**.336**	.094	.048
Ministral-3-14B	**.014**	.000	.000	**.572**	.088	.004
Llama-3.2-3B	**.026**	.000	.000	**.708**	.140	.048
Llama-3.1-8B	**.070**	.002	.030	**.614**	.184	.252
Llama-3.2-11B-Vision	**.080**	.006	.044	**.516**	.138	.284

Figure 8 tests how the comparison among RGS, BiPO, and RePS changes as the diagnostic threshold increases. Larger κ imposes a stricter margin requirement, so the diagnostic rate decreases or stays constant; the informative features are the gaps and crossings between methods. For example, RePS has a higher point estimate than RGS at the upper end of the displayed threshold range for Qwen3-8B. The bands for these three formal methods are unstudentized simultaneous 95% bands over the complete stored discrete grid $\kappa \in [0, 4.5]$, not over a continuum. The plotted window is $[0, 2]$, so the full-grid bands are conservative for the displayed range. The 10,000 bootstrap replicates resample sources within each fixed pair, share draws across directions and methods, and re-estimate s_m; the five pairs remain fixed and are macro-averaged equally. The intervals capture Validation-source uncertainty conditional on the frozen response families, five pair identities, single training seed, and selected steering configurations. Appendix G.3 adds the two pairwise diagnostic probes to the same curves.

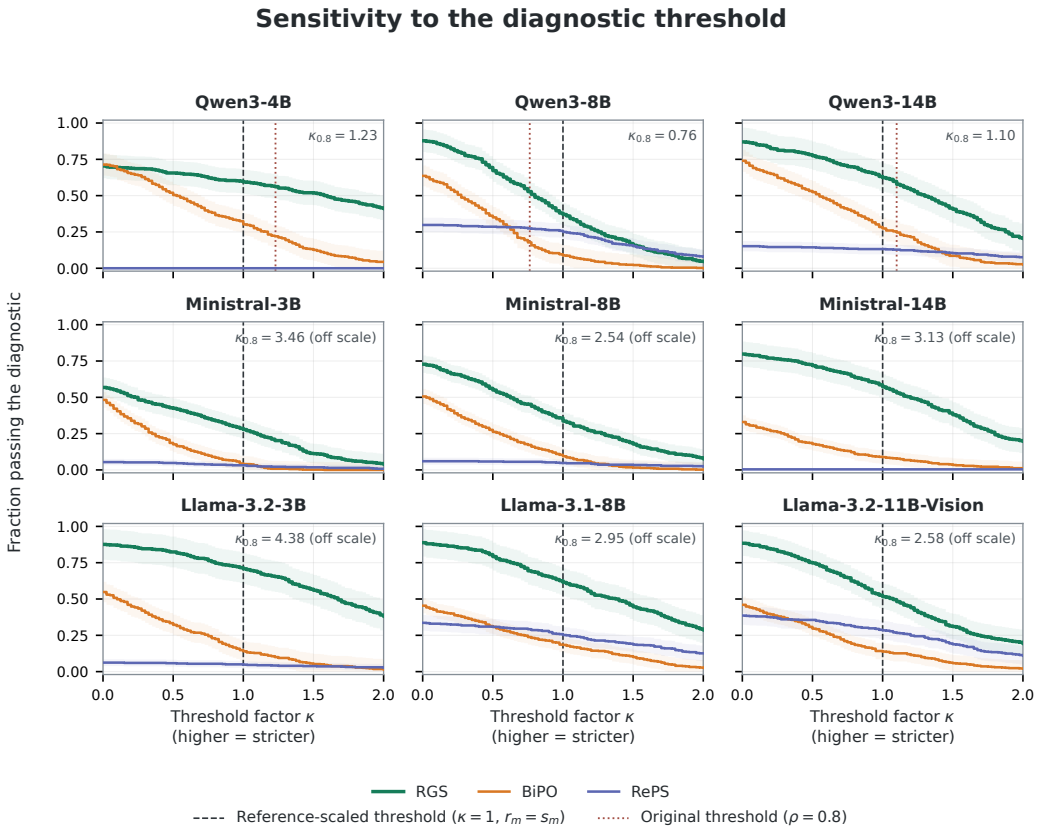


Figure 8: Threshold sensitivity of the RGS, BiPO, and RePS comparison across nine fixed models. Larger κ in $r_m(\kappa) = \kappa s_m$ imposes a stricter diagnostic threshold. The black dashed line marks the reference-scaled threshold at $\kappa = 1$; the red dotted line marks the original threshold $\rho = 0.8$, at $\kappa = 0.8/s_m$, and is omitted when it falls outside the displayed $[0, 2]$ range. Shading gives simultaneous 95% source-cluster bootstrap bands for the three methods, computed over the stored $[0, 4.5]$ grid.

These Validation geometries at the selected steering configurations diagnose teacher-forced $(\Delta_I, \Delta_A, \Delta_N)$ distributions on the five training-pair identities. Appendix E studies strict-Unseen generated behavior and Judge outcomes along sampled endpoint chords.

Appendix D.7 shows that full RGS, with steering configurations selected on five training-pair identities, achieves higher BPC, Order-preserving rate, and Target-value Induction Rate than RePS and BiPO on the 40 untrained and unselected strict-Unseen pair recombinations across all nine models. These generation metrics do not by themselves establish intended-support uplift relative to the unsteered response or satisfaction of the teacher-forced control specification. The matched Qwen3-4B results trace the tested objective-to-behavior chain: pairwise data-fit leaves absolute movement underidentified; the zero-anchored LogSumExp competitor maximum supplies reference pressure against suppression-based solutions; in the seed-42 alternative-only Δ_A design, the enhance term changes mean intended movement from negative to positive and leakage improves alternative-behavior control conditional on the enhance term. In the supplementary singleton-A endpoint factorial, the bundle changes the Separate-minus-Signed BPC effect from negative to positive. Under the complete $0/A/N$ objective, separate endpoints instead have a BPC advantage both without and with the bundle; the bundle improves the associated intended-uplift and quality tradeoff, with little change in that BPC advantage. Under the targeted Qwen3-4B challenge settings, the A and N members further protect their respective competitor-specific margins.

G COMPLETE DIAGNOSTIC FIGURES

G.1 COMPLETE PAIR-LEVEL INTERPOLATION TRAJECTORIES

Figures 9– 17 report the complete set of 40 strict-Unseen pair-level trajectories for every evaluated model, using the definitions in Appendix E. Dotted lines mark each method's unsteered baseline; annotations count adjacent reversals in RGS/BiPO/RePS order.

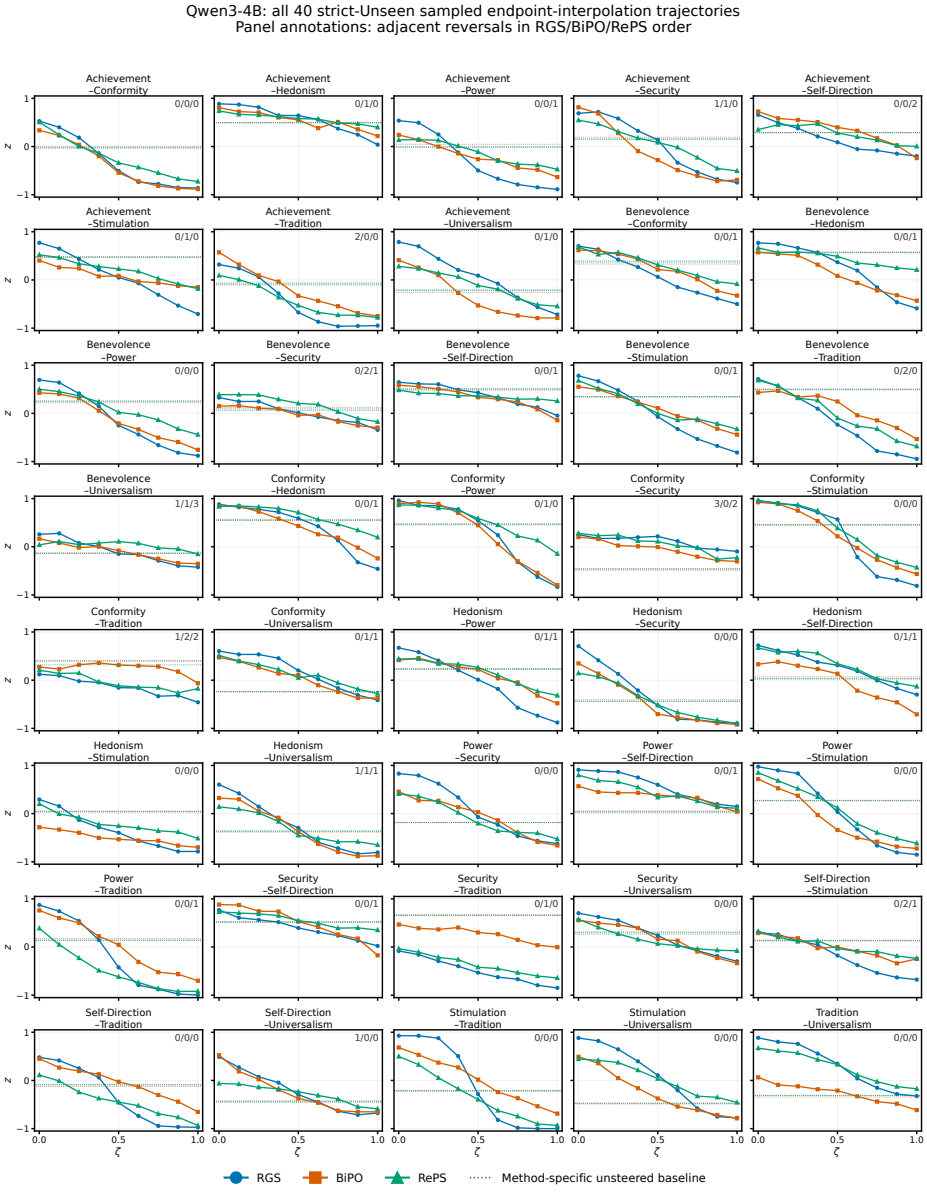


Figure 9: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Qwen3-4B. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

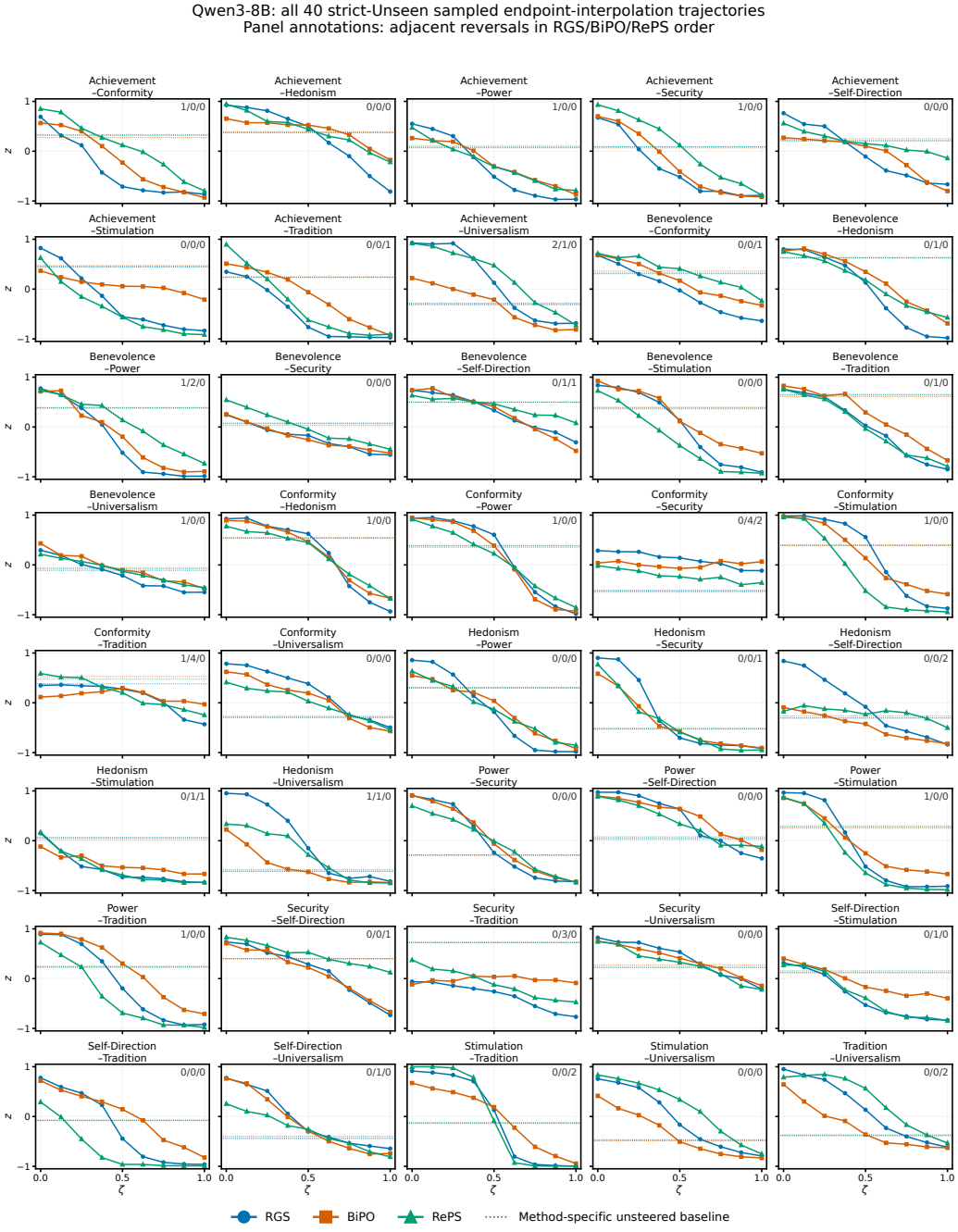


Figure 10: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Qwen3-8B. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

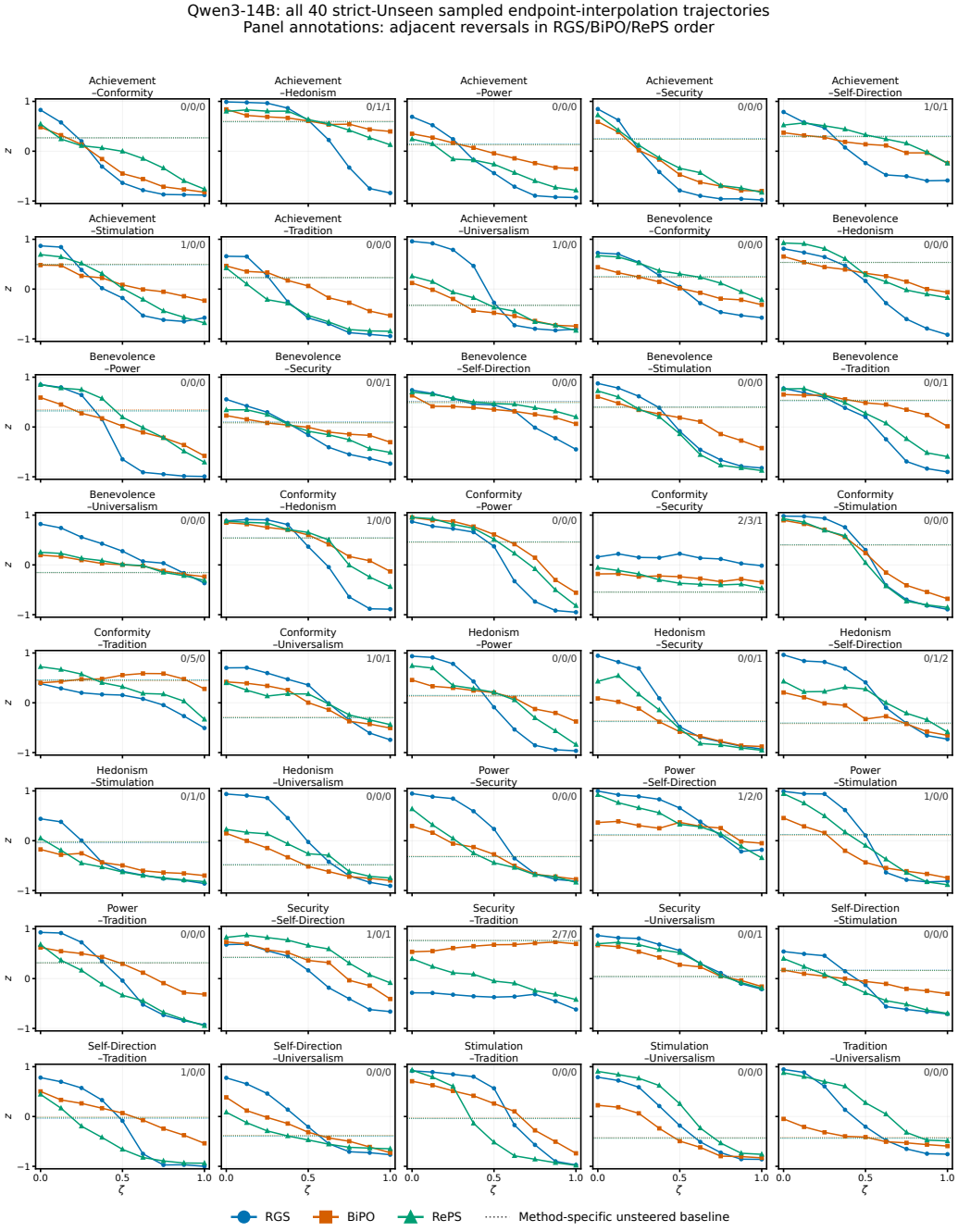


Figure 11: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Qwen3-14B. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

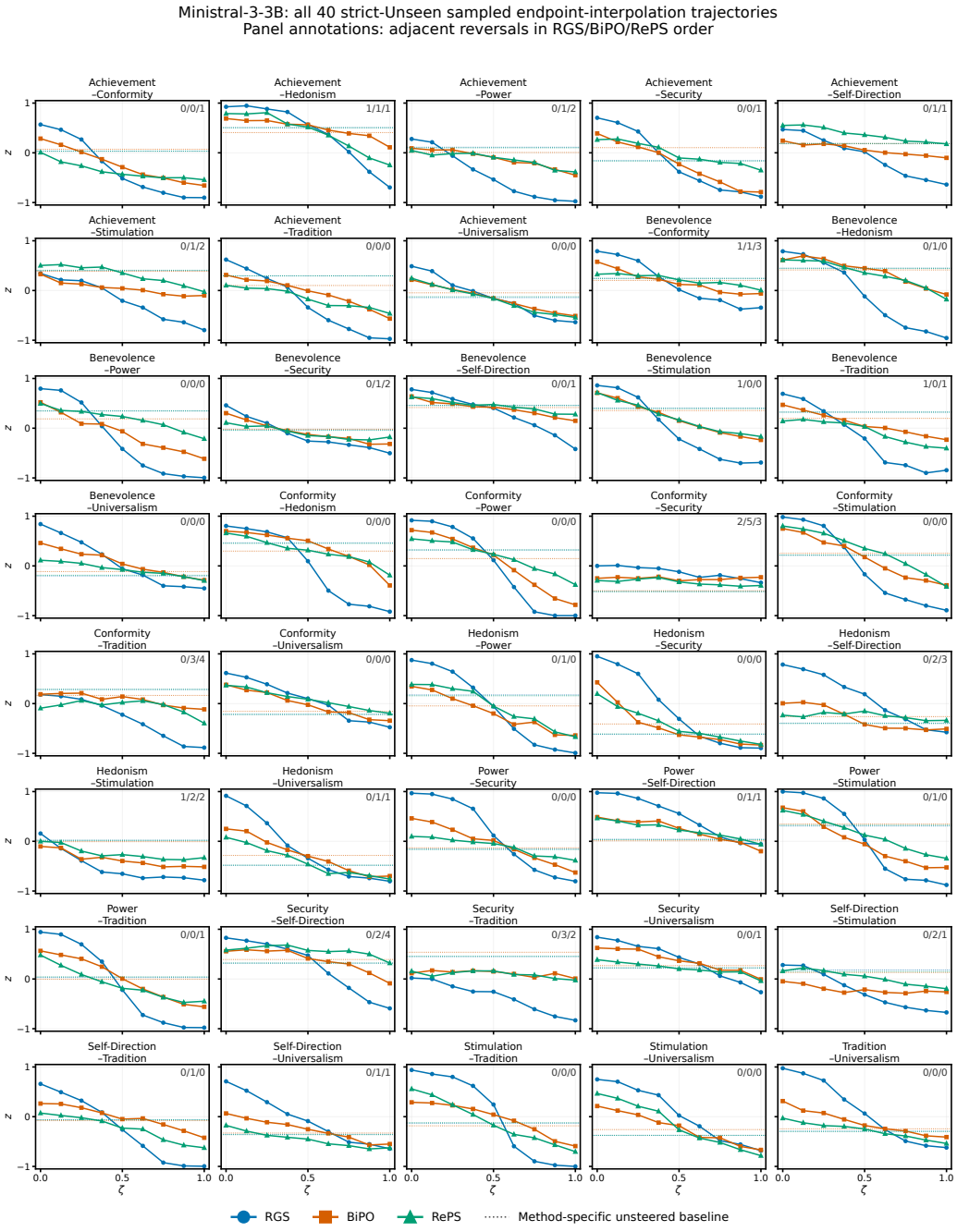


Figure 12: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Ministral-3-3B. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

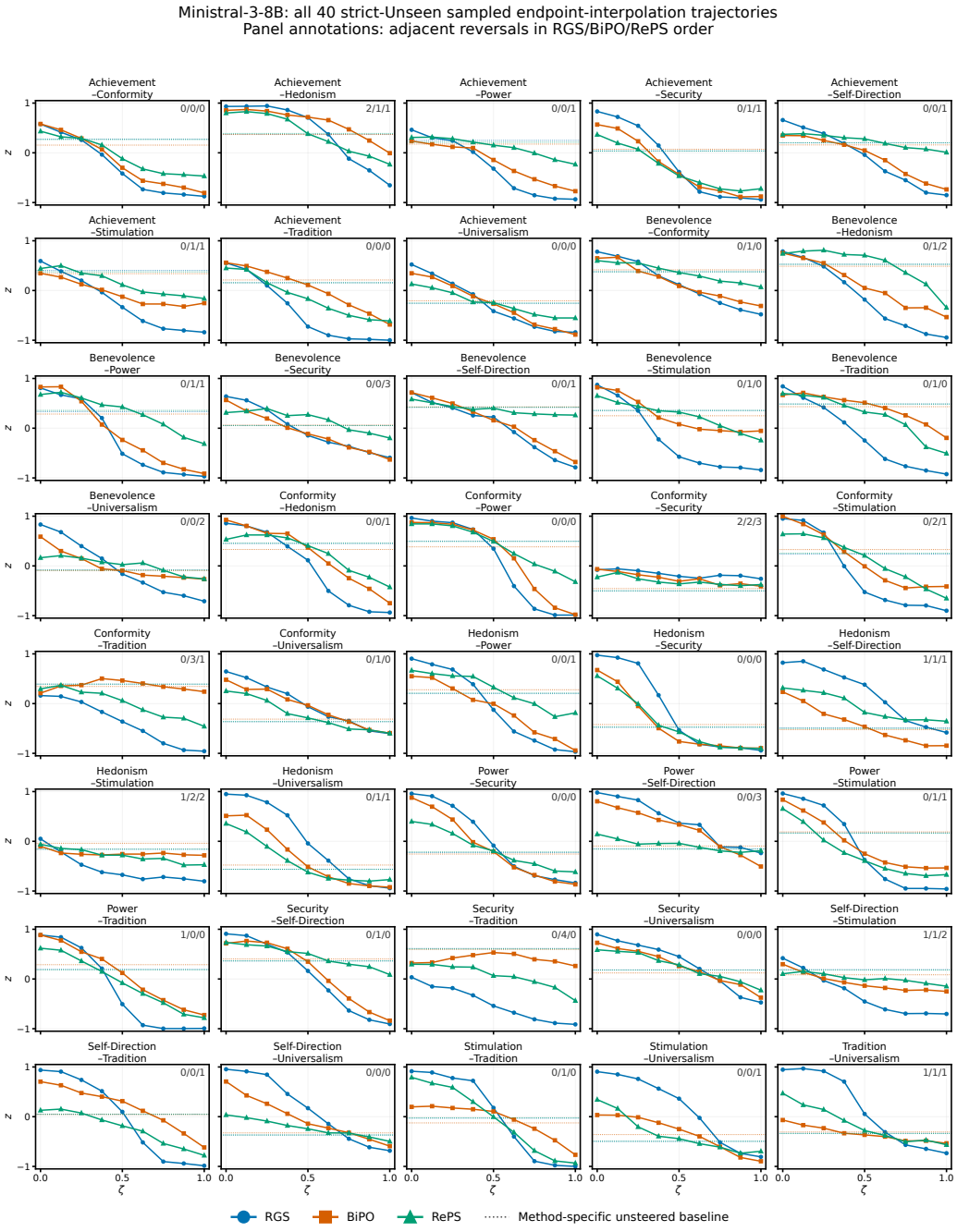


Figure 13: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Ministral-3-8B. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

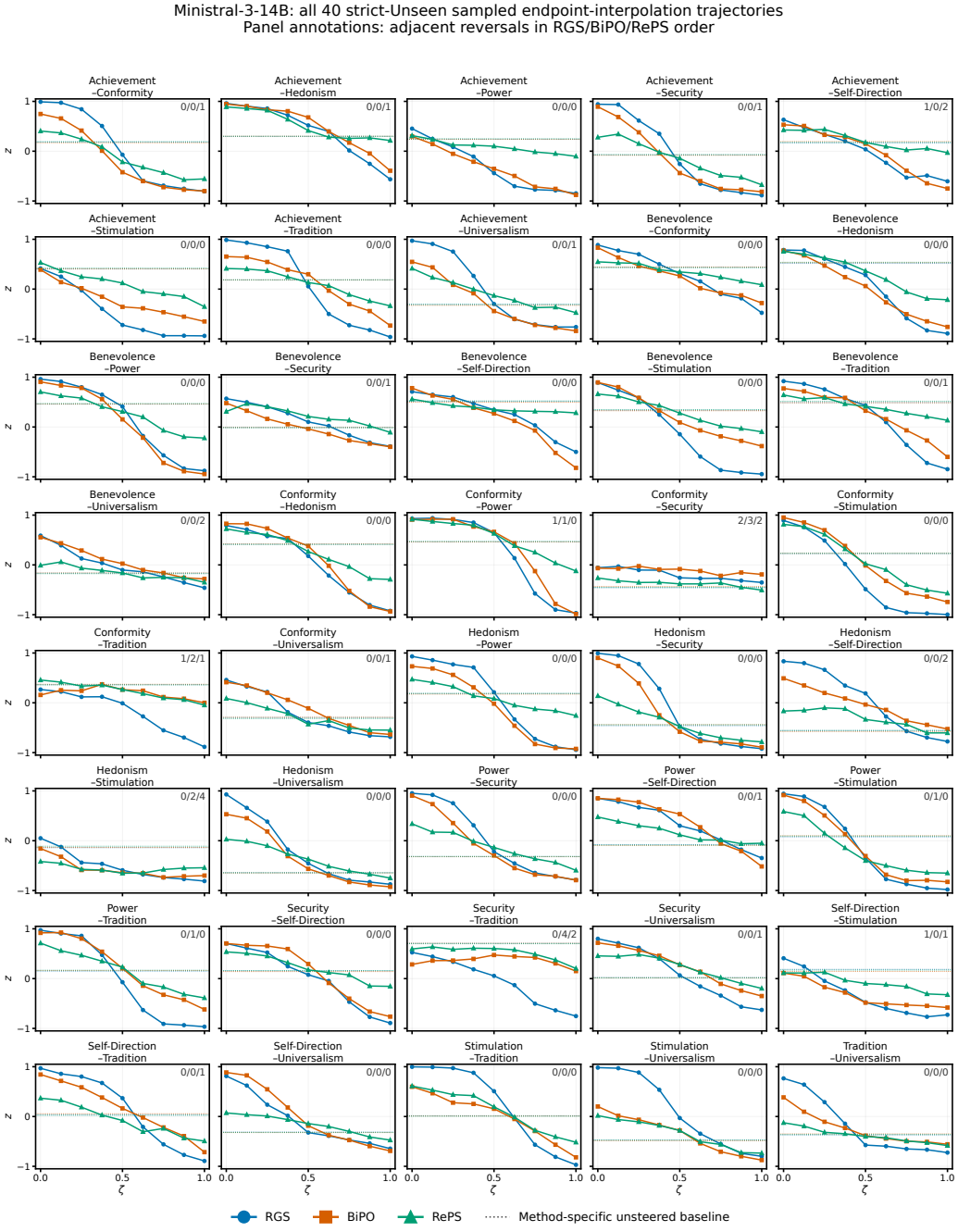


Figure 14: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Ministral-3-14B. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

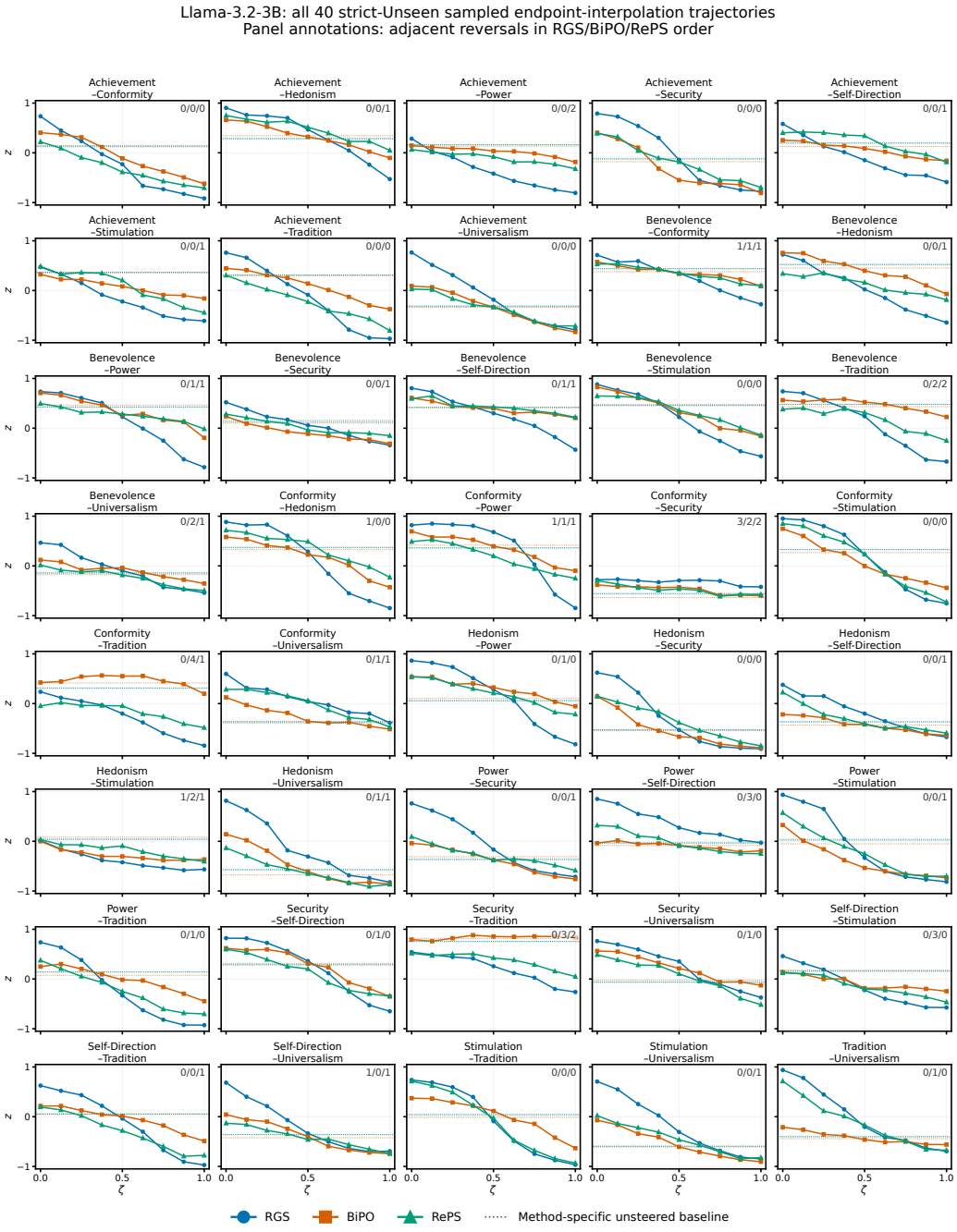


Figure 15: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Llama-3.2-3B. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

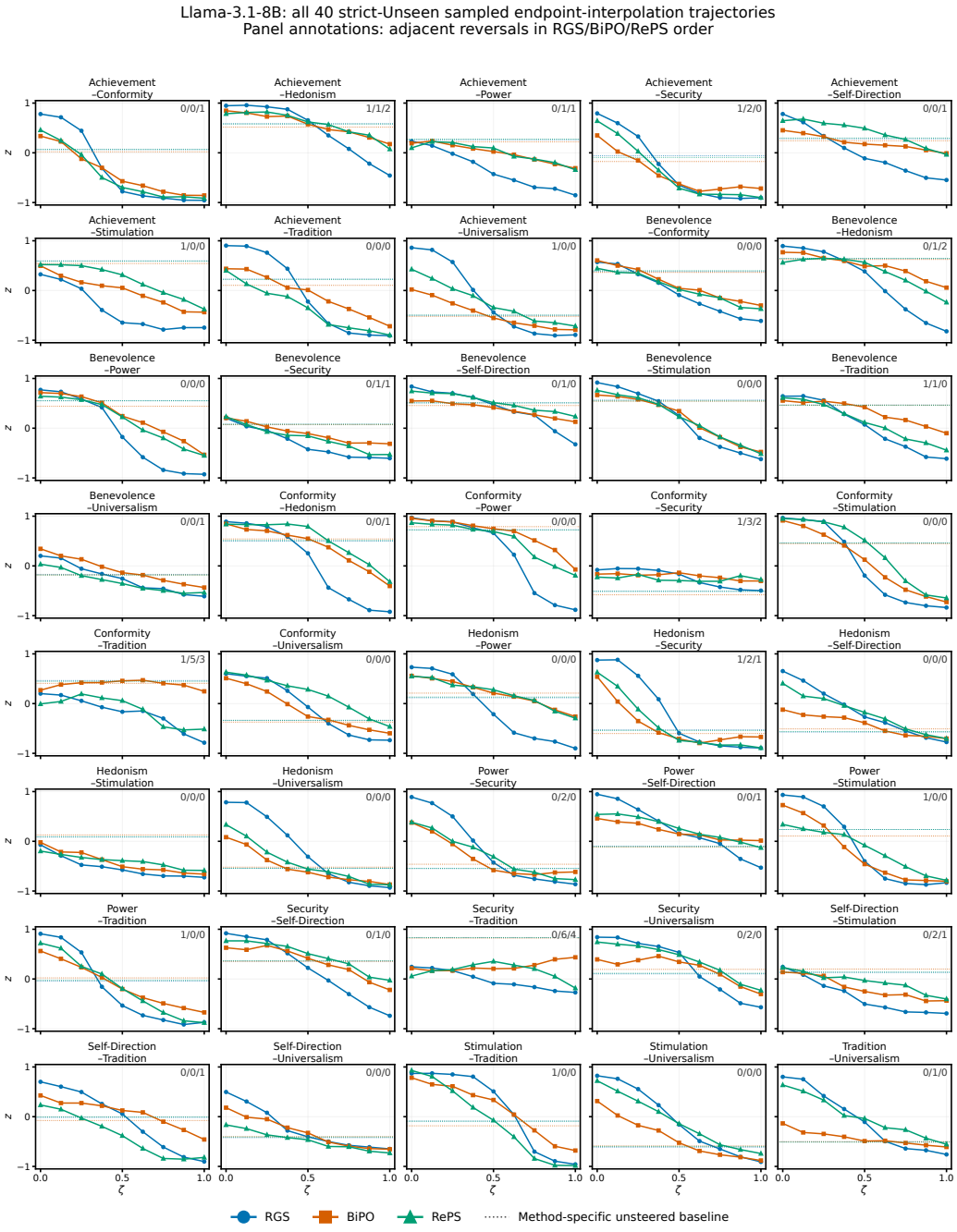


Figure 16: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Llama-3.1-8B. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

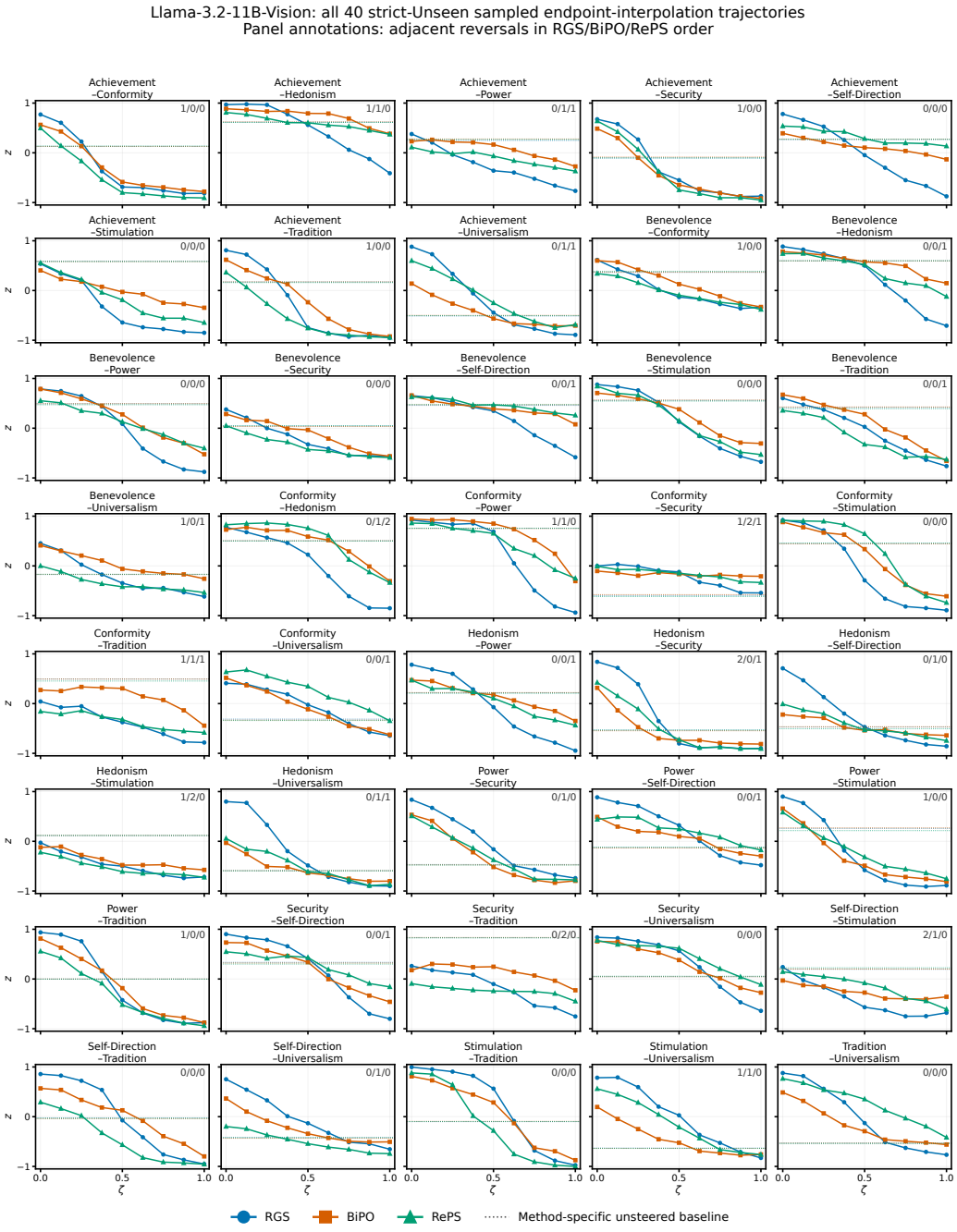


Figure 17: All 40 strict-Unseen sampled endpoint-interpolation trajectories for Llama-3.2-11B-Vision. Panel annotations give adjacent-reversal counts in RGS/BiPO/RePS order.

G.2 ADDITIONAL MODEL GEOMETRY VIEWS

Figures 18– 21 show the remaining eight models. All panels use the same absolute threshold $\rho = 0.8$. Each model uses separate pooled Δ_N point-color scales for the formal-method and pairwise-probe rows. Colors are therefore comparable only within the same model and row, not across rows or models; triangular colorbar tips mark clipping at the corresponding 98th-percentile absolute displacement.

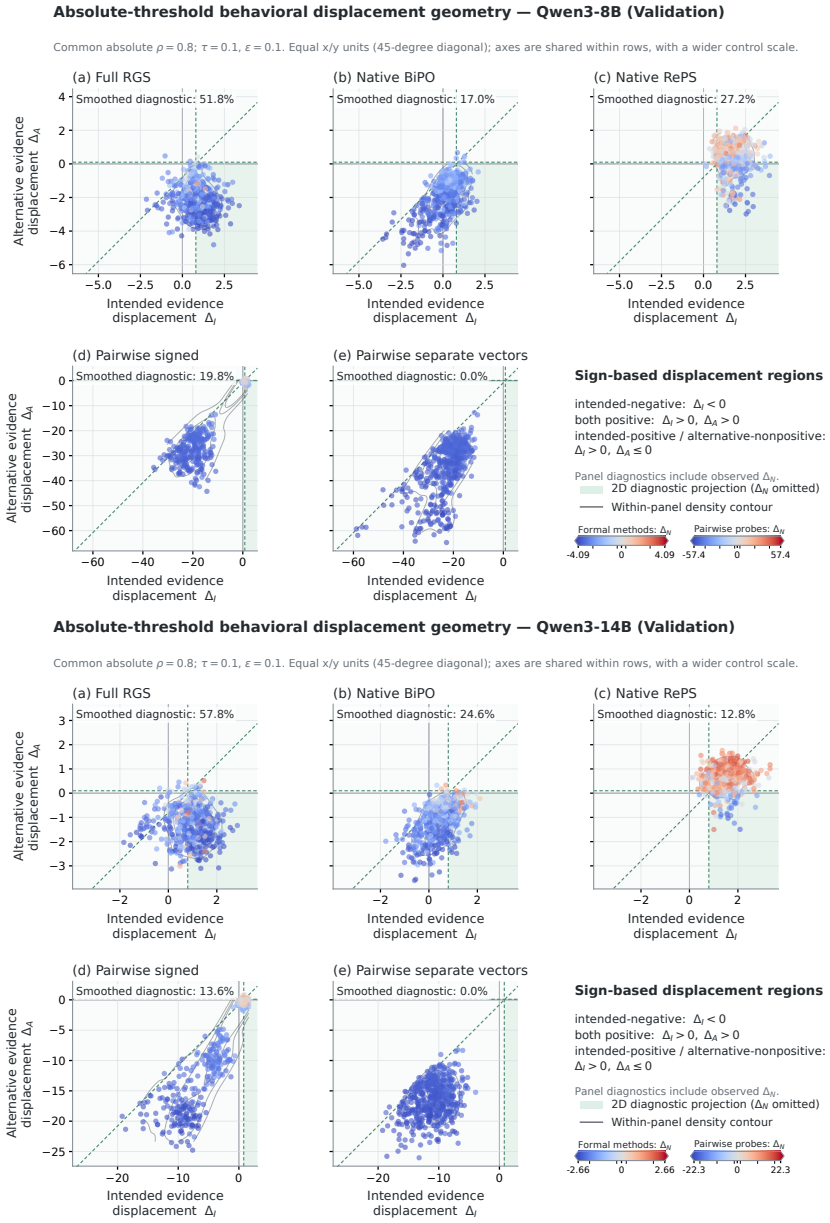


Figure 18: Behavioral geometry at selected steering configurations under the fixed threshold $\rho = 0.8$ for Qwen3-8B (upper) and Qwen3-14B (lower). Records and plotting conventions follow Figure 6.

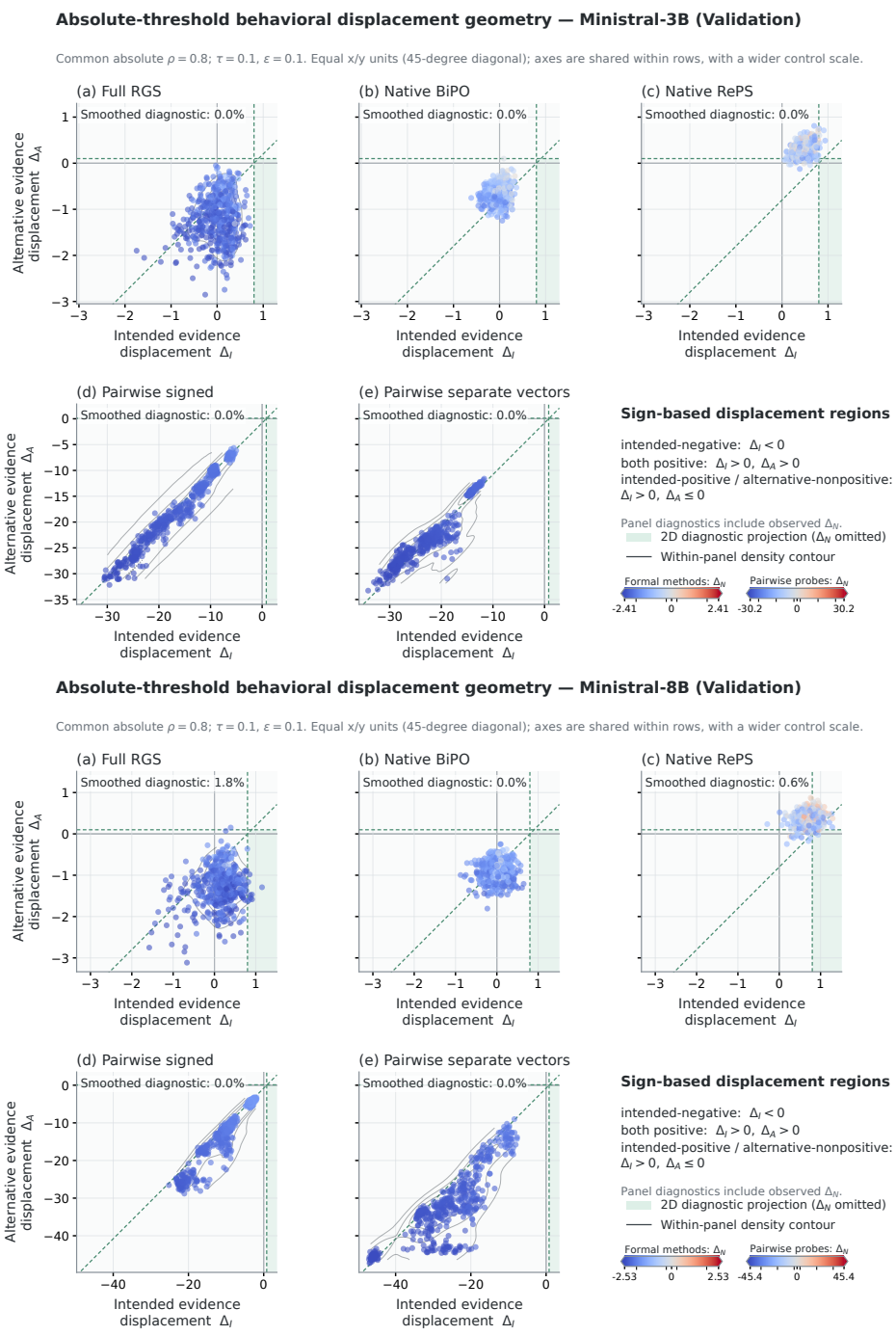


Figure 19: Behavioral geometry at selected steering configurations under the fixed threshold $\rho = 0.8$ for Ministral-3-3B (upper) and Ministral-3-8B (lower). Plotting conventions follow Figure 6.

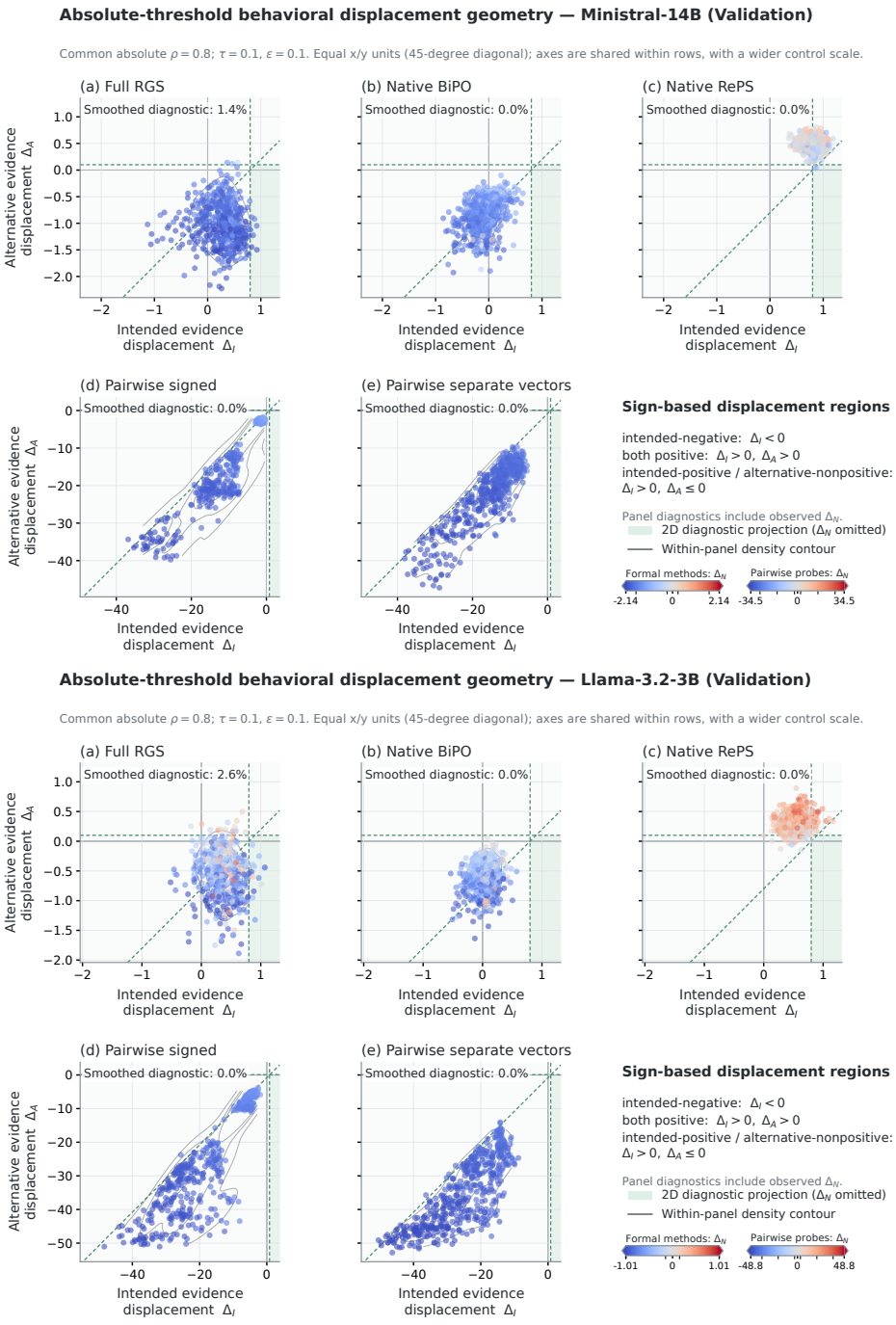


Figure 20: Behavioral geometry at selected steering configurations under the fixed threshold $\rho = 0.8$ for Ministral-3-14B (upper) and Llama-3.2-3B (lower). Plotting conventions follow Figure 6.

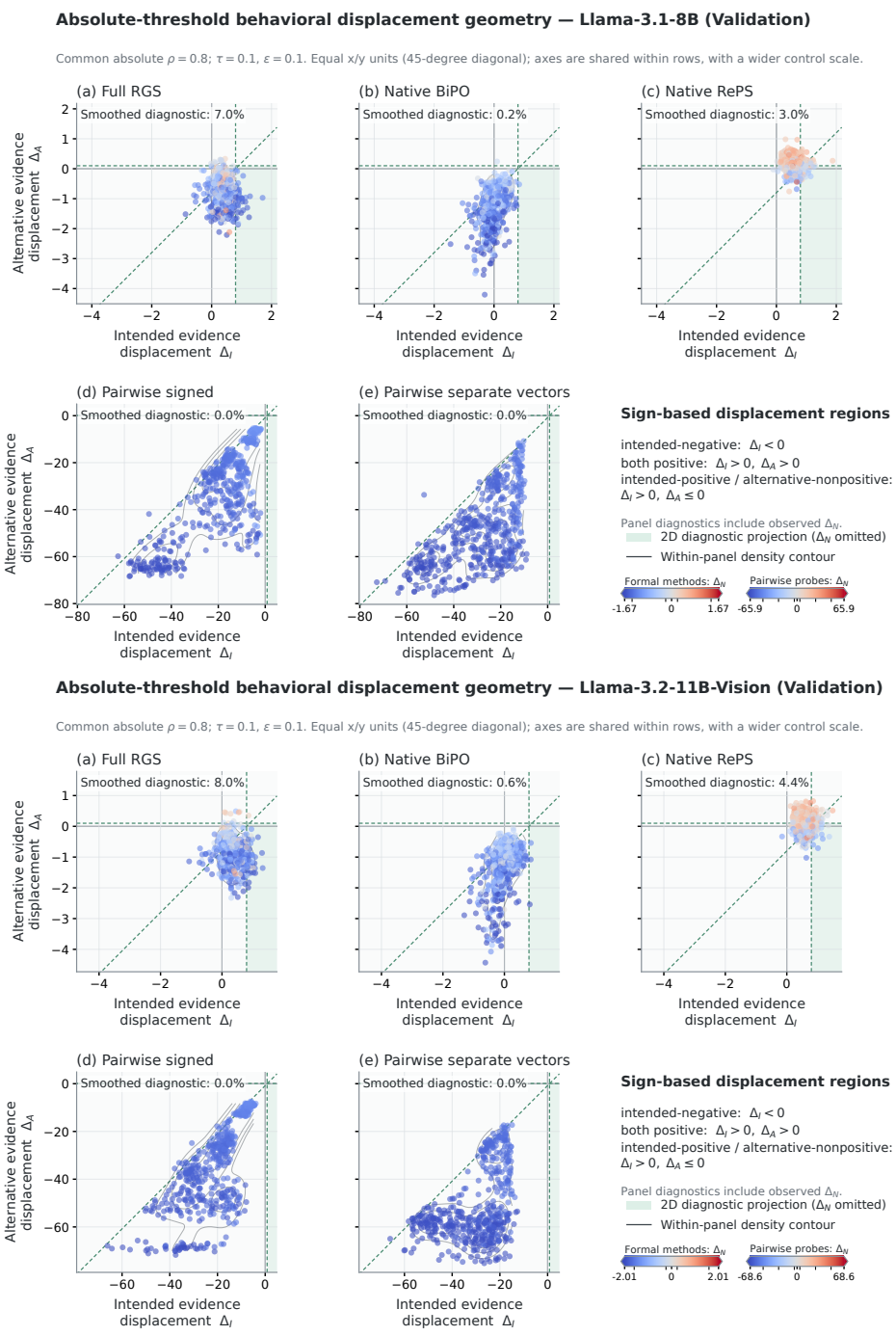


Figure 21: Behavioral geometry at selected steering configurations under the fixed threshold $\rho = 0.8$ for Llama-3.1-8B (upper) and Llama-3.2-11B-Vision (lower). Plotting conventions follow Figure 6.

G.3 THRESHOLD-SENSITIVITY CURVES WITH PAIRWISE PROBES

Figure 22 extends Figure 8 with the pairwise signed and separate-vector probes. These probes use the corresponding RGS layer and strength. The diagnostic, reference scales, and bootstrap procedure follow Appendix F.4.3.

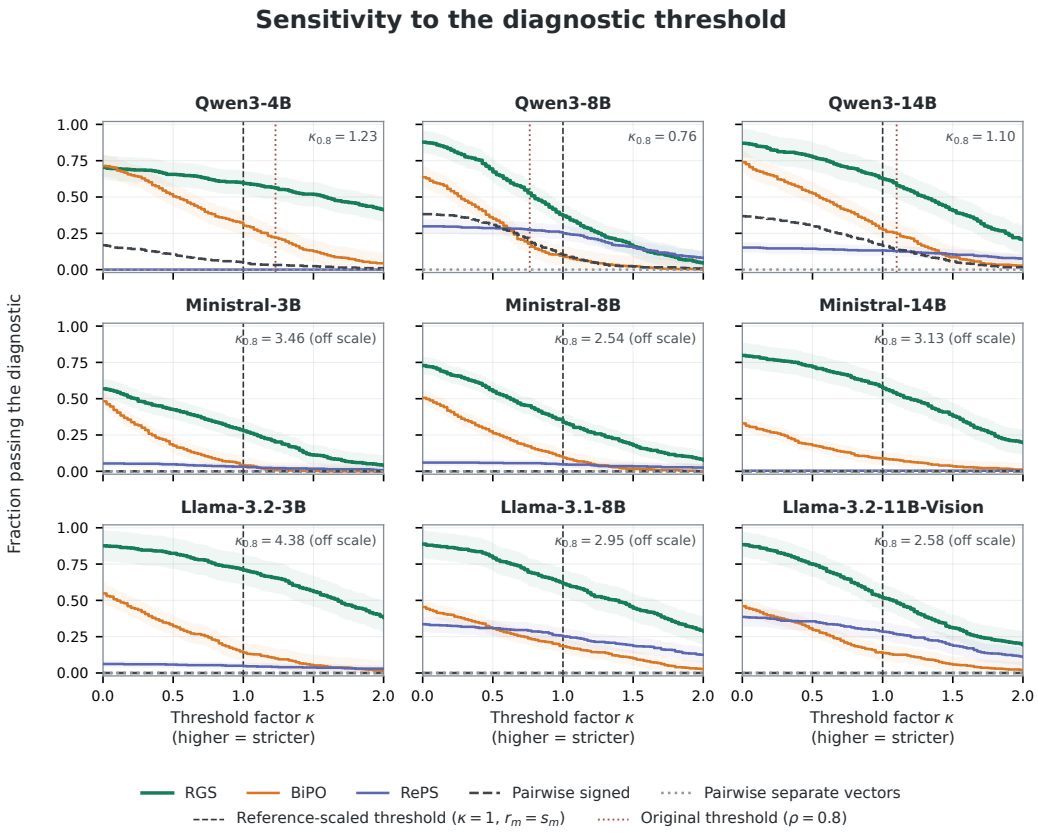


Figure 22: Threshold-sensitivity curves for the three formal methods and two pairwise diagnostic probes across all nine models. Larger κ gives a stricter threshold $r_m(\kappa) = \kappa s_m$. The black dashed line marks $\kappa = 1$; the red dotted line marks the original $\rho = 0.8$ threshold at $\kappa = 0.8/s_m$ and is omitted outside the displayed $[0, 2]$ range. Shading gives simultaneous 95% source-cluster bootstrap bands for RGS, BiPO, and RePS over the stored $[0, 4.5]$ grid; the two probes are shown without bands.